

Search for the Production of a Standard Model Higgs Boson in Association with Top-Quarks and Decaying into a Pair of Bottom-Quarks with 13 TeV ATLAS Data

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium (Dr. rer. nat.)

im Fach Physik

eingereicht an der

Mathematisch-Naturwissenschaftlichen Fakultät

der Humboldt-Universität zu Berlin

von

M.Sc. Nedaa Alexandra Asbah

Präsident der Humboldt-Universität zu Berlin

Prof. Dr. Sabine Kunst

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät

Prof. Dr. Elmar Kulke

Gutachter:

1. PD Dr. Judith Katzy
2. Prof. Dr. Thomas Lohse
3. PD Dr. Hannes Jung

Tag der mündlichen Prüfung: 23.05.2018

Erklärung

Ich erkläre, dass ich die Dissertation selbständig und nur unter Verwendung der von mir gemäß §7 Abs. 3 der Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät, veröffentlicht im Amtlichen Mitteilungsblatt der Humboldt-Universität zu Berlin Nr. 126/2014 am 18.11.2014 angegebenen Hilfsmittel angefertigt habe.

I declare that I have produced this doctor's thesis independently using only the tools I have specified, in accordance with section 7 para. 3 of the Faculty of Mathematics and Natural Sciences PhD regulations, published in the Official Gazette of Humboldt-Universität zu Berlin (Amtliches Mitteilungsblatt) no. 126/2014 on 18/11/2014.

Hamburg, 05.03.2018

Nedaa Alexandra Asbah

IN MEMORY OF MY AUNT

SAWSAN ASBAH
(03.11.1962-21.11.2015)

ACKNOWLEDGMENTS

This thesis would not have been possible without the presence of all those who surrounded me, both professionally and personally.

I am deeply grateful to my supervisor, Judith Katzy, for her endless support and confidence in my abilities in the past four years as a PhD student at DESY. I appreciate all her time and effort, particularly during the write up of this thesis.

I would like to thank all my current and previous colleagues and friends at DESY and CERN; Andrea, Cécile, Chris, Claire, Daniel, Ralph, Jelena, John, Loïc, Mirko, Snežana Spyros, Stefan, Timothée, Valerie, Valerio, and special thanks to Paul for reviewing this thesis and to Georges for the endless help and always answering my questions. You all made my time in Hamburg and CERN great!

Thanks to all my colleagues in the FTK group, especially Stefan, Jeremy, Rui, and Jordan for the great work and experience.

I owe my deepest gratitude to my family; my dear parents Mona and Bader, sisters Nimara and Malvina, my uncle Khader, and my cousins Nathalie, Maroupie, Anwar, and Karmen for their continuous support, and love.

Finally thanks to all my friends all over the world who accompanied me in my life up to now, particularly my friend Chris for making my stay in Geneva as amazing as it was, and my friends Maria and Suleiman for the great weekends in Hamburg. Above all, special thanks to Jeff for his endless love, support, patience, and confidence in me.

ABSTRACT

This thesis presents the search for the Standard Model Higgs boson produced in association with a pair of top-quarks ($t\bar{t}H$). The analysis uses a 36.1 fb^{-1} dataset of proton-proton collisions at a center-of-mass energy of $\sqrt{s} = 13 \text{ TeV}$ collected with the ATLAS detector, at the Large Hadron Collider during 2015 and 2016. The analysis presented here searches for the $t\bar{t}H$ production in the $H \rightarrow b\bar{b}$ decay mode. The selected events contain either one or two leptons from the decay of the top-quark pair. In order to improve the sensitivity of the search, events are split in regions according to the number of jets and how likely these events are to contain b -jets. Methods based on multivariate techniques were developed and applied in the signal-enriched regions to discriminate $t\bar{t}H$ events against background events being dominated by top pair production with additional b -jets, $t\bar{t} + \geq 1b$. Detailed studies presented here have been performed to estimate the background from $t\bar{t} + \geq 1b$. Moreover, misidentification of leptons causes background events in the analysis samples which are estimated using a data-driven technique based on the Matrix Method. All analysis regions are combined in a statistical model using a profile likelihood fit to constrain the background predictions and reduce the systematic uncertainties. For a Higgs boson mass of 125 GeV , the ratio of the measured $t\bar{t}H$ signal cross section to the Standard Model expectations, $\mu_{t\bar{t}H}$, is found to be

$$\mu_{t\bar{t}H} = 0.84^{+0.64}_{-0.61}.$$

An excess of events over the expected Standard Model background is found with an observed (expected) significance of 1.4 (1.6) standard deviations. A $t\bar{t}H$ signal strength larger than 2.0 is excluded at the 95% confidence level.

ZUSAMMENFASSUNG

Die vorliegende Arbeit beschreibt die Suche nach der Produktion des Standardmodell Higgs-Bosons in Assoziation mit einem Top-Antitop-Quarkpaar ($t\bar{t}H$). Der verwendete Datensatz basiert auf einer integrierten Luminosität von 36.1 fb^{-1} , aufgenommen mit dem ATLAS Detektor am Large Hadron Collider in den Jahren 2015 und 2016 bei einer Schwerpunktsenergie von $\sqrt{s} = 13 \text{ TeV}$. Die Analyse wurde für den Zerfall des Higgs-Bosons in zwei b -Quarks ($H \rightarrow b\bar{b}$) konstruiert und die selektierten Ereignisse enthalten entweder ein oder zwei Leptonen vom Zerfall des Top-Antitop-Quarkpaares. Die Sensitivität der Analyse wurde erhöht, indem die Ereignisse in unterschiedliche Regionen unterteilt wurden, basierend auf der Anzahl der Jets sowie der Wahrscheinlichkeit b -Jets zu enthalten. Methoden basierend auf multivariaten Analysetechniken wurden entwickelt, um $t\bar{t}H$ Signalereignisse vom Untergrund zu separieren, der von der Produktion von Top-Antitop-Quarkpaaren mit zusätzlichen b -Jets dominiert wird. Detaillierte Studien wurden durchgeführt um den dominierenden $t\bar{t} + \geq 1b$ Untergrund abzuschätzen. Des Weiteren wurden Untergrundereignisse, die die Selektion aufgrund der Misidentifikation von Leptonen passieren, mit der auf Daten basierenden Matrix Methode abgeschätzt. Alle in der Analyse verwendeten Regionen wurden in einem Profile-Likelihood-Fit kombiniert, um die Vorhersagen des Untergrunds einzuschränken und die systematischen Unsicherheiten zu reduzieren. Das Verhältnis des gemessenen $t\bar{t}H$ Wirkungsquerschnitts zur Standardmodell-Vorhersage, $\mu_{t\bar{t}H}$, beträgt

$$\mu_{t\bar{t}H} = 0.84^{+0.64}_{-0.61}$$

bei einer Higgs-Boson Masse von 125 GeV . Ein Überschuss an Ereignissen über dem erwarteten Standardmodell-Untergrund wurde mit einer beobachteten (erwarteten) Signifikanz von 1.4 (1.6) Standardabweichungen gemessen. Die Daten schließen $t\bar{t}H$ Signalstärken von mehr als 2.0 mit einem Konfidenzniveau von 95% aus.

Table of Contents

ACKNOWLEDGMENTS	i
ABSTRACT	iii
1 INTRODUCTION	1
2 THE STANDARD MODEL OF PARTICLE PHYSICS	5
2.1 Brief Description of the Standard Model	5
2.2 Electroweak Theory	8
2.3 Symmetry Breaking: The Higgs Mechanism	10
2.4 The Higgs Boson	14
2.4.1 Production Mechanisms of the Higgs Boson	15
2.4.2 Higgs Boson Decays	17
2.4.3 Properties of the Higgs Boson	19
2.5 The Top-Quark	19
2.5.1 Top-Quark Pair Production	20
2.5.2 Decay of the Top-Quark	21
2.6 Direct Measurement of the Top Higgs Yukawa Coupling	23
3 SIMULATION OF PARTICLE INTERACTIONS	25
3.1 Quantum Chromodynamics	26
3.2 The Factorization Theorem: PDFs and the DGLAP Equations	28
3.3 Matrix Element	30
3.4 Parton Shower	31
3.5 Hadronization	33
3.6 Underlying Event	34
3.7 Monte Carlo Generators	34
3.8 ATLAS Simulation	36
4 THE LHC AND THE ATLAS DETECTOR	39
4.1 The Large Hadron Collider	39
4.2 Luminosity and Pileup	41
4.3 The ATLAS Detector	42
4.3.1 The Inner Detector	43
4.3.2 Calorimeter	47
4.3.3 The Muon Spectrometer	51
4.3.4 The Trigger and Data Acquisition	53
4.3.5 Fast TracKer	55
4.3.6 Luminosity Measurement	57

5	DEFINITION OF PHYSICS OBJECTS	59
5.1	Tracks and Primary Vertices	59
5.2	Leptons	61
5.2.1	Electrons	61
5.2.2	Muons	67
5.3	Jets	68
5.3.1	Jet Reconstruction	69
5.3.2	Jet Calibration	71
5.3.3	Jet Energy Scale Uncertainty	74
5.3.4	Jet Energy Resolution	75
5.3.5	Jet Vertex Tagger	76
5.4	b -tagging	77
5.4.1	b -tagging Algorithms	79
5.4.2	b -tagging Calibration	81
5.5	Missing Transverse Energy	82
6	ESTIMATION OF FAKE AND NON-PROMPT LEPTONS	85
6.1	Processes for Faking Electrons and Muons	85
6.1.1	Sources of Fake and Non-prompt Electrons	86
6.1.2	Sources of Fake and Non-prompt Muons	86
6.2	Modeling Fake Events	88
6.3	The Vanilla Matrix Method	88
6.3.1	Fake and Real Efficiencies	91
7	SEARCH FOR THE PRODUCTION OF A STANDARD MODEL HIGGS BOSON IN ASSOCIATION WITH TOP-QUARKS AND DECAYING INTO A PAIR OF BOTTOM-QUARKS	95
7.1	Measurement of $t\bar{t}H$ in the $(H \rightarrow b\bar{b})$ Decay Mode	95
7.2	Search for $t\bar{t}H(H \rightarrow b\bar{b})$ at 8 TeV	97
7.3	Data and Simulation Samples	98
7.3.1	Data Taking	98
7.3.2	Triggers	99
7.3.3	Simulated Samples	100
7.4	Object Selection	106
7.5	Event Selection and Categorization	108
7.5.1	Event Selection of Recorded Data	108
7.5.2	Event Categorization at Particle Level	109
7.5.3	Event Categorization at Reconstruction Level	110
7.6	Multivariate Analysis	117
7.6.1	Boosted Decision Trees	117
7.6.2	MVA-based Reconstruction of the $t\bar{t}H$ Final State	120
7.6.3	Discrimination between Signal and Background	124
7.7	Background Estimation	129
7.7.1	$t\bar{t}$ +jets Background	129

7.7.2	$t\bar{t} + \geq 1b$ Background	131
7.7.3	Fake and Non-prompt Lepton Background	142
7.7.4	Other Backgrounds	158
7.8	Kinematic Distributions in the Analysis Regions	158
7.9	Systematic Uncertainties	165
7.9.1	Experimental Uncertainties	165
7.9.2	Uncertainties Related to the Background Estimation	166
7.9.3	Uncertainties on the Signal Modeling	170
7.9.4	Summary of Systematic Uncertainties	170
7.10	Statistical Analysis and Results	173
7.10.1	The Profile Likelihood Fit	173
7.10.2	The Fit Model	175
7.10.3	Expected Performance	178
7.10.4	Fit to Data and Results	185
7.10.5	Setting Limits	197
8	CONCLUSION AND OUTLOOK	201
A	ADDITIONAL MATERIAL	205
A.1	The Boosted Category	205
A.2	Region Definition in the dilepton channel	205
A.3	$t\bar{t}$ +HF modeling	208
A.4	Event Yields	213
A.5	Additional plots	215
A.6	Setting limits	217
	REFERENCES	219

Chapter 1

INTRODUCTION

The Standard Model (SM) of elementary particles and their interactions has so far provided a remarkably accurate description of particle physics. The recent discovery of the SM-like Higgs particle, by both the ATLAS and CMS experiments at the Large Hadron Collider (LHC) in 2012 [1, 2], ended the 40-year hunt for this particle. The Higgs boson was the last missing piece of the SM [3–5]. The discovery confirms the success of the proposed theory about the existence of an associated Higgs field that describes electroweak symmetry breaking as a mechanism to generate massive vector bosons [6–10], in addition to fermion masses through Yukawa coupling.

Even though the Higgs boson has been discovered by ATLAS and CMS in excess of 5σ , and initial measurements confirm that it is compatible with the description of the SM [11–15], there is still uncertainty surrounding its attributes. It is essential to precisely measure the properties of the Higgs boson as any potential disagreement could hint at fascinating new phenomena. For example, the fermion Yukawa couplings are not well measured. Since the top-quark is the heaviest fundamental particle in the SM, its coupling to the Higgs boson is expected to be the strongest. A precise measurement of this coupling is a stringent test of the SM, and any deviation could be very sensitive to physics beyond the Standard Model (BSM) [16].

The ATLAS and CMS collaborations made a first attempt to extract the top-quark Yukawa coupling in proton-proton collisions from the inclusive Higgs boson production cross section or decay, using data collected at the LHC with a center of mass energy of 7 and 8 TeV during 2010–2012, referred to as Run 1 of the LHC. This coupling determination relies largely on the gluon-gluon fusion production mode and on the decay mode to photons, which both depend on loop contributions with a top-quark. A combined signal yield relative to the SM predictions is measured to be equal to 1.09 ± 0.11 [17]. In order to measure the Higgs coupling to the top-quark directly without any assumptions on loop processes, the measurement of the production of a SM Higgs boson in association with a pair of top-quarks ($t\bar{t}H$) is needed.

The $t\bar{t}H$ process is a rare production mode at the LHC, with only 1% of the total Higgs boson production cross section [18]. However, its cross section at $\sqrt{s} = 13$ TeV, increased

by almost a factor of four compared to its cross section at $\sqrt{s} = 8$ TeV, providing a unique opportunity to perform this measurement. The Higgs boson decays into various pairs of SM particle. Of these, the decay to two b -quarks is predicted to have a branching fraction about 58% [18], the largest Higgs boson decay mode.

The $t\bar{t}H(H \rightarrow b\bar{b})$ channel suffers from the large backgrounds from the production of top-quark pairs with additional jets ($t\bar{t}$ +jets), especially when the associated jets contain b - or c -hadrons. Furthermore, due to the presence of b -quarks from top decays, combinatorial ambiguity arising from the many jets originating from b -quarks (b -jets) in the final state, makes it challenging to find the two b -jets originating from the Higgs boson and to identify the signal events.

This thesis presents a search for the $t\bar{t}H$ production using 36.1 fb^{-1} of pp collision data at $\sqrt{s} = 13$ TeV. This data was collected from the start of Run 2 of the LHC during 2015 and 2016. The analysis targets Higgs boson decays to b -quarks, but all the decay modes may contribute to the signal. Events are required to have one or two leptons from the decay of the top-quark pair and exclusive analysis categories are defined based on the number of leptons, the number of jets and the value of a b -tagging discriminant that provides a measure of how likely a jet contains a b -hadron; i.e. is originating from a b -quark.

Multivariate techniques based on Boosted Decision Trees (BDT) are used to reconstruct the $t\bar{t}H$ signal and to distinguish it from the large $t\bar{t}$ +jets background. Crucial for this measurement is an adequate and precise estimation of the dominant background arising from $t\bar{t}$ production with additional b -jets. Therefore, the modeling has been studied extensively and a combination of the most up to date theoretical predictions and a sophisticated statistical analysis have been developed to constrain the large background uncertainties. The signal-rich categories are analyzed together with the signal-depleted ones in a combined likelihood fit that simultaneously determines the event yields for the signal while constraining the overall background model within the assigned systematic uncertainties.

The thesis is structured as follows. The important role of the Higgs mechanism in the SM and the coupling of fermions to the Higgs field is described in Chapter 2. It also details the Higgs boson and top-quark production and decay modes at the LHC, in particular the $b\bar{b}$ decay mode which is of interest in this thesis.

Chapter 3 summarises the Monte Carlo (MC) simulations that serve as the theoretical predictions of the signal and background processes.

The objects from the decay of the complex $t\bar{t}H$ final state are measured by multiple sub-systems within the ATLAS detector, introduced in Chapter 4. The particle reconstruction and identification of the measured objects are detailed in Chapter 5.

Despite the sophisticated reconstruction and identification algorithms, misidentification of reconstructed objects might still happen, causing background events in the analysis sample. Chapter 6 presents a data-driven technique based on the Matrix Method for estimating this background.

The main topic of this thesis, the search for $t\bar{t}H(H \rightarrow b\bar{b})$, is detailed in Chapter 7. The first part of this chapter summarizes the selection criteria applied to events and physics objects, describes the event categorization, and details the multivariate analysis techniques used to reconstruct the $t\bar{t}H$ signal events and to separate them from the dominate $t\bar{t}$ +jets background. The second part details the background estimation and the assigned systematic uncertainties with emphasise on the background arising from the $t\bar{t} + b\bar{b}$ process. The last part of this chapter, introduces the fit model, and the results.

An overall summary and conclusion of these studies is given in Chapter 8, alongside an outlook for future round of this analysis with the full 13 TeV dataset.

Chapter 2

THE STANDARD MODEL OF PARTICLE PHYSICS

This chapter presents an overview of the Standard Model (SM) with a brief summary of the elementary particles and their fundamental interactions. Particular emphasis is dedicated to describe how particles acquire their mass through their interaction with the Higgs field. The Higgs mechanism of electroweak symmetry breaking, predicts the existence of a massive scalar particle, the Higgs boson. The second part of this chapter discusses the production and decay mechanisms of the Higgs boson as well as the top-quark.

2.1 Brief Description of the Standard Model

The SM, which was formulated in the 1960s and 70s [3, 4, 6, 8, 10, 19], represents our current understanding of elementary particles and their interactions. The particle content of the SM is shown in Figure 2.1. The elementary particles consist of two types of particles, fermions and bosons. Leptons and quarks together form the fermions and are arranged in three families. They are spin-1/2 particles which obey Fermi-Dirac statistics. Particles interact with each other through the four fundamental forces, the strong, the electromagnetic, the weak and the gravitational force. The gravitational force is not yet included in the SM. However, it is by far the weakest and its effect is assumed negligible in the interactions of elementary particles. Fermions interact through mediators, referred to as gauge bosons, that act as force carriers. The force carriers are integer spin particles that obey Bose-Einstein statistics. Stable matter in our universe is composed of electrons, the up- and down-quarks that form the first family, or sometimes referred to as the first generation, of particles in the SM. The particles that belong to the two other generations have the same quantum numbers as the particles in the first generation but differ by their masses.

Each particle is subject to interactions that are determined by their quantum numbers. Quarks are the fundamental constituents of hadrons and they have a unique attribute called color charge that dictates their interaction through the strong force mediated by massless spin-1 gluons (g). Color charge comes in three different types: *red*, *blue*, and *green*. The gluon exists in eight different states and carries a combination of color and

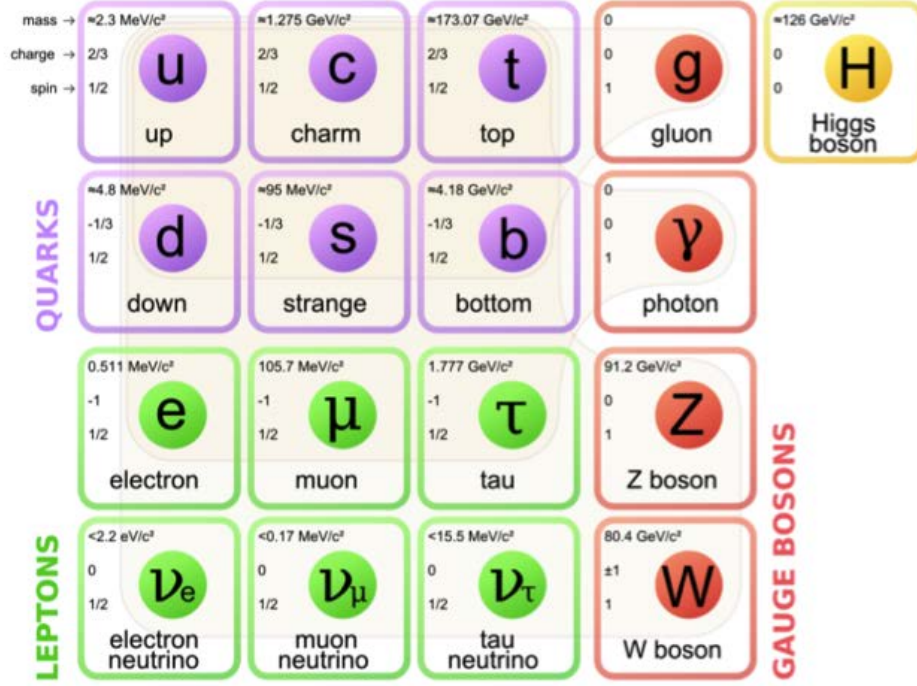


Figure 2.1: The standard model (SM) of elementary particles consisting of three generations of quarks, leptons, and neutrinos as well as five force carrying bosons [20]. For each particle, the mass, spin and charge is given.

anti-color charge. The up-type quarks (up (u), charm (c), top (t)) carry an electric charge (e) of $+2/3e$, while the down-type quarks (down (d), strange (s), bottom (b)) carry an electric charge of $-1/3e$. The charged leptons (electron (e), muon (μ), tau (τ)) have an integer charge ($-1e$). Electrically charged particles interact through the electromagnetic force mediated by massless neutral spin-1 photons (γ). Each charged lepton is associated to a neutral lepton (electron-, muon-, tau-neutrino (ν_e, ν_μ, ν_τ)). All the mentioned particles interact via the weak force, mediated by the electrically charged W^\pm or the neutral Z vector bosons. The W^\pm and the Z bosons have considerable masses and spin of one.

The SM is a quantum field theory (QFT), implying that its fundamental objects are quantum fields which are defined at all points in space-time. It is based upon the Lagrangian formalism and the fundamental notion of symmetries. It describes the interaction among the components of matter, fermions, through the exchange of force mediators, bosons.

Fermions are described as spin-1/2 Dirac fields that satisfy the following Lagrangian:

$$\mathcal{L} = \bar{\psi}(i\gamma^\mu\partial_\mu - m)\psi, \quad (2.1)$$

where ψ is the fermion field, γ^μ are the Dirac matrices, and m is the fermion mass. The SM Lagrangian \mathcal{L}_{SM} can be written based on two parts: the quantum chromodynamics (QCD) Lagrangian, \mathcal{L}_{QCD} describing the strong interactions, and the electroweak Lagrangian, \mathcal{L}_{EW} describing the electromagnetic and weak interactions:

$$\mathcal{L}_{\text{SM}} = \mathcal{L}_{\text{QCD}} + \mathcal{L}_{\text{EW}}. \quad (2.2)$$

Following Noether's theorem [21], for every differentiable symmetry generated by a local action, there is a corresponding conserved current.

The SM is founded on the gauge symmetry of the group

$$G_{\text{SM}} = SU(3)_C \times SU(2)_L \times U(1)_Y, \quad (2.3)$$

where (C) stands for color, (L) represents that the symmetry applies to only left-handed fields, and ($Y \equiv 2(Q - T_3)$) stands for the weak hypercharge¹. The first term of the gauge group ($SU(3)_C$) is a non-Abelian symmetry² group that refers to the color symmetry of QCD. The second group ($SU(2)_L \times U(1)_Y$) is associated with the transformations of the weak isospin and hypercharge of the leptons and quarks and is related to the conservation of the corresponding quantities. These symmetries dictate the SM's internal generators³ which are eight for $SU(3)_C$, three for $SU(2)_L$, and one for $U(1)_Y$.

The $SU(2)_L \times U(1)_Y$ symmetry in the SM requires the EW mediators to be massless. However, experimental measurements have shown that the three electroweak gauge bosons (W^\pm, Z) are massive, which require an explicit mass term in the Lagrangian that violates the gauge invariance. The Brout-Englert-Higgs mechanism [6, 8, 10] introduces a spontaneous $SU(2)_L \times U(1)_Y$ symmetry breaking that solves the inconsistency among the SM theory and the measurements. The elementary particles acquire their masses through their

1. T_3 represents the projection of the weak isospin along the z-axis and Q stands for the electric charge

2. Non-Abelian here means that the symmetry operations do not commute $[a, b] \neq 0$, $a, b \in G$, where G stands for a group.

3. Simple unitary group $SU(N)$ has $N^2 - 1$ generators

interactions with the Higgs field, which will be detailed in Section 2.3 of this chapter. The strength of the interaction of the particle determines its acquired mass which is proportional to the Higgs field. This mechanism predicts a scalar particle, the Higgs boson, whose mass is a free parameter of the theory.

2.2 Electroweak Theory

The theory of electroweak interactions (EW) [22] explains the decay of muons, neutrons, and top-quarks. It describes how the weak charged current and electromagnetic processes are invariant under the weak hyper-charge $U(1)$ and the weak isospin $SU(2)$ transformations. Therefore, this theory is invariant under the transformations of the gauge group $SU(2)_L \times U(1)_Y$.

The first part of the symmetry group $SU(2)_L$ introduces the weak isospin, T , where the generators of the group are the weak isospin operators: $\hat{T} = \frac{\sigma_i}{2}$ ($i = 1, 2, 3$), where σ_i are the Pauli matrices.

The left- and right-handed fermion fields are grouped in the EW theory in the following:

$$\psi_L = \frac{1}{2}(1 - \gamma^5)\psi, \quad \psi_R = \frac{1}{2}(1 + \gamma^5)\psi, \quad (2.4)$$

where $\frac{1}{2}(1 \pm \gamma^5)$ are the chirality operators and $\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3$. In this description, the left-handed fermions form weak isospin doublets ($I = \frac{1}{2}$) and the right-handed fermions are isospin singlets ($I = 0$):

$$\begin{pmatrix} u \\ d \end{pmatrix}_L, \begin{pmatrix} \nu_e \\ e \end{pmatrix}_L, \quad \begin{pmatrix} c \\ s \end{pmatrix}_L, \begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix}_L, \quad \begin{pmatrix} t \\ b \end{pmatrix}_L, \begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}_L \quad (2.5)$$

$$u_R, d_R, e_R, \quad c_R, s_R, \mu_R, \quad t_R, b_R, \tau_R. \quad (2.6)$$

The second part of the symmetry group $U(1)_Y$, introduces the hypercharge Y . The electric charge of a particle is associated with the hypercharge Y and the third component of the weak isospin T_3 via the Gell-Mann Nishijima formula:

$$\hat{Q} = \frac{\hat{Y}}{2} + \hat{T}_3 \quad (2.7)$$

A covariant derivative is introduced in Equation 2.1 in order to respect the local invariance under both symmetry groups as:

$$D_\mu \equiv \partial_\mu - ig\vec{T} \cdot \vec{W}_\mu - ig' \frac{Y}{2} B_\mu, \quad (2.8)$$

in which g and g' are the coupling constants of both gauge groups $SU(2)_L$ and $U(1)_Y$, respectively. The gauge fields of the symmetry groups are represented by \vec{W}_μ and B_μ .

The mentioned gauge fields require a kinetic term in the Lagrangian, which has the following form:

$$\mathcal{L}_{\text{gauge}} = -\frac{1}{4} W_{\mu\nu}^i W_i^{\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu}, \quad (2.9)$$

in which $i = 1, 2, 3$, $W_{\mu\nu}^i$ and $B_{\mu\nu}$ are field tensors for the $SU(2)_L$ and $U(1)_Y$ gauge groups, represented as:

$$W_{\mu\nu}^i \equiv \partial_\mu W_\nu^i - \partial_\nu W_\mu^i + g \varepsilon_{jk}^i W_\mu^j W_\nu^k \quad (2.10)$$

$$B_{\mu\nu} \equiv \partial_\mu B_\nu - \partial_\nu B_\mu, \quad (2.11)$$

where ε_{jk}^i is the antisymmetric Levi-Civita tensor.

The electroweak Lagrangian will be

$$\mathcal{L}_{\text{EW}} = \sum_{f=l,q} \bar{f} i \gamma^\mu D_\mu f + \mathcal{L}_{\text{gauge}}. \quad (2.12)$$

The Lagrangian density does not contain terms related to the gauge boson and fermion masses. However, this is not in agreement with experiments that confirmed the existence of massive fermions and electroweak mediators with masses of $m_{W^\pm} = 80.4$ GeV and $m_{Z^0} = 91.2$ GeV [20]. By spontaneously breaking the symmetry with the Higgs mechanism, the gauge bosons and the fermions acquire their masses through the interaction with the Higgs field, known as the Yukawa Interaction.

2.3 Symmetry Breaking: The Higgs Mechanism

A proposed solution to accommodate massive gauge fields is the so-called *Higgs mechanism* which causes Spontaneous Symmetry-Breaking (SSB), where the symmetry group $SU(2)_L \times U(1)_Y$ breaks down to $U(1)_{EM}$. To successfully attain the SSB, an additional isospin doublet of complex scalar fields, known as the Higgs field, is introduced:

$$\phi = \begin{pmatrix} \phi^\dagger \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi^1 + i\phi^2 \\ \phi^3 + i\phi^4 \end{pmatrix}. \quad (2.13)$$

The Lagrangian for this field is

$$\mathcal{L}_H = (D_\mu \phi)^\dagger (D^\mu \phi) - V(\phi), \quad (2.14)$$

which consists of a kinetic term and a Higgs potential ($V(\phi)$)

$$V(\phi) = \mu^2 \phi^\dagger \phi + \lambda (\phi^\dagger \phi)^2 = \mu^2 \phi^2 + \lambda \phi^4. \quad (2.15)$$

The first term in Equation 2.15 can be associated with the mass of the field, while the second term stands for the self-interaction of the field. The minima of the potential (ϕ_0) can be identified with the vacuum expectation value (v) of the Higgs field. The vacuum expectation value is defined as the absolute value of the field at the minimum of the potential⁴. However, the parameter (λ) of the potential needs to be positive since the case of $\lambda < 0$ is unphysical. The parameter (μ) can be chosen freely; for $\mu^2 > 0$ the potential V assumes a unique minimum at $\phi_0 = 0$, leading to a symmetric ground state under $SU(2)$. On the other hand, for $\mu^2 < 0$, the shape of the potential is modified as illustrated in Figure 2.2, in which V assumes a non-trivial minimum $\phi_0^2 = -\frac{\mu^2}{2\lambda} \equiv \frac{v^2}{2}$, and the choice of the physical vacuum state spontaneously breaks the symmetry of the Lagrangian.

The Goldstone theorem implies that massless scalars, referred to as Goldstone bosons, appear when a continuous symmetry is broken [23]. A gauge field can absorb the massless scalars as a longitudinal polarization component, and as a consequence acquire mass. The minimum of the potential is chosen in a way that the Higgs field that acquires the vacuum

4. This is only true at leading order.

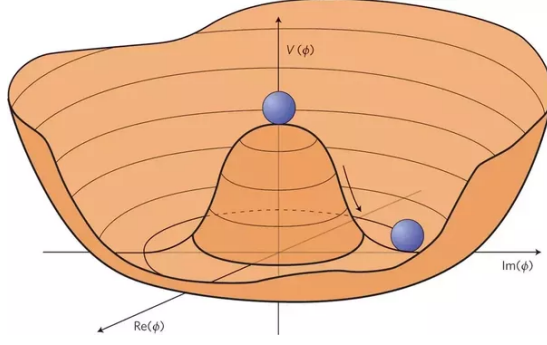


Figure 2.2: The shape of the Higgs potential $V(\phi) = \mu^2\phi^2 + \lambda\phi^4$. The vacuum state is randomly chosen from infinite number of choices when falling into the vacuum state which leads to spontaneous symmetry breaking.

expectation value is the one with zero electric charge:

$$\langle 0|\phi|0\rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}. \quad (2.16)$$

An expansion around the reference minimum,

$$\phi'(x) = \frac{e^{i\vec{\sigma}\cdot\vec{\theta}(x)/v}}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H(x) \end{pmatrix}, \quad (2.17)$$

where $H(x)$ denotes the massive scalar Higgs field, and θ stands for the three fields which will be absorbed by the gauge fields. Then, the Lagrangian of the Higgs field becomes:

$$\mathcal{L}_H = (D_\mu H)^\dagger (D^\mu H) - \frac{1}{2}(-2\mu^2)H^2 - \lambda v H^3 - \frac{1}{4}\lambda H^4. \quad (2.18)$$

The second term in Equation 2.18 corresponds to the tree-level mass term of the $H(x)$ field, which is computed from the Higgs Lagrangian represented in Equation 2.14 to be

$$m_H = \sqrt{-2\mu^2} = \sqrt{2\lambda}v. \quad (2.19)$$

Since λ is not predicted, the theory does not predict m_H , and it needs to be determined experimentally.

From the same Higgs Lagrangian defined in Equation 2.14, the electroweak boson

masses can be obtained as:

$$\begin{aligned}
\left| \left(-ig \frac{\vec{\sigma}}{2} \vec{W}_\mu - i \frac{g'}{2} B_\mu \right) \Phi \right|^2 &= \frac{1}{8} \left| \begin{pmatrix} gW_\mu^3 + g'B_\mu & g(W_\mu^1 - iW_\mu^2) \\ g(W_\mu^1 + iW_\mu^2) & -gW_\mu^3 + g'B_\mu \end{pmatrix} \begin{pmatrix} 0 \\ v \end{pmatrix} \right|^2 \\
&= \frac{1}{8} v^2 g^2 \left[(W_\mu^1)^2 + (W_\mu^2)^2 \right] + \frac{1}{8} v^2 (g'B_\mu - gW_\mu^3)(g'B_\mu - gW_\mu^3) \\
&= \left(\frac{1}{2} v g \right)^2 W_\mu^+ W_\mu^- + \frac{1}{8} v^2 \begin{pmatrix} W_\mu^3 & B_\mu \end{pmatrix} \begin{pmatrix} g^2 & -gg' \\ -gg' & g'^2 \end{pmatrix} \begin{pmatrix} W_\mu^3 \\ B_\mu \end{pmatrix},
\end{aligned} \tag{2.20}$$

where the charged fields are defined as $W^\pm = (W^1 \mp iW^2)/\sqrt{2}$. The mass eigenstates can be acquired by diagonalizing the mass matrix, and expressed in terms of W_μ^3 and B_μ :

$$\begin{aligned}
\frac{1}{8} v^2 \left[g^2 (W_\mu^3)^2 - 2gg'W_\mu^3 B_\mu + g'^2 B_\mu^2 \right] &= \frac{1}{8} v^2 \left[gW_\mu^3 - g'B_\mu \right]^2 \\
&\quad + 0 \left[g'W_\mu^3 + gB_\mu \right]^2 \\
&= \frac{1}{2} \left(v \frac{\sqrt{g^2 + g'^2}}{2} \right)^2 Z_\mu^2 \\
&\quad + 0 \cdot A_\mu^2,
\end{aligned} \tag{2.21}$$

where the neutral physical fields (the Z boson and the photon fields) are defined as

$$Z_\mu = \frac{gW_\mu^3 - g'B_\mu}{\sqrt{g^2 + g'^2}} \quad \text{and} \quad A_\mu = \frac{g'W_\mu^3 + gB_\mu}{\sqrt{g^2 + g'^2}}. \tag{2.22}$$

By introducing the *weak mixing angle* θ_W

$$\cos \theta_W = \frac{g'}{\sqrt{g^2 + g'^2}}, \quad \sin \theta_W = \frac{g}{\sqrt{g^2 + g'^2}}, \tag{2.23}$$

the neutral fields can be rewritten as

$$Z_\mu = -B_\mu \sin \theta_W + W_\mu^3 \cos \theta_W \quad \text{and} \quad A_\mu = B_\mu \cos \theta_W + W_\mu^3 \sin \theta_W. \tag{2.24}$$

The masses of the gauge bosons are deduced from the quadratic terms in the field defined in Equations 2.20 and 2.21, giving $M_W = \frac{gv}{2}$ and $M_Z = \frac{\sqrt{g^2+g'^2}v}{2}$, while the photon remains massless. The masses of the gauge bosons are related to each other through the weak mixing angle:

$$\frac{M_W}{M_Z} = \cos \theta_W. \quad (2.25)$$

As briefly mentioned in Section 2.2, fermion masses are also generated through the spontaneous breaking of the $SU(2)_L \times U(1)_Y$ gauge symmetry by introducing a Yukawa term that describes the interaction among the fermion and the Higgs fields. The interaction between the Higgs and the fermion fields in the form the Yukawa Lagrangian is:

$$\mathcal{L}_{\text{Yukawa}} = \sum_{f=l,q} y_f [\bar{f}_L \phi f_R + \bar{f}_R \bar{\phi} f_L], \quad (2.26)$$

where y_f are the matrices containing the Yukawa coupling constants between the fermions and the Higgs boson. The Yukawa Lagrangian is gauge invariant since the combinations $\bar{f}_L \phi f_R$ and $\bar{f}_R \bar{\phi} f_L$ are $SU(2)_L$ singlets.

The matrices y_f can be diagonalized in order to get the eigenvalues of the Yukawa couplings using unitary transformations that will redefine the fermion fields. In the leptonic sector this transformation has no effect due to the absence of right-handed neutrinos. However, in the quark sector, the rotation to the mass eigenstate basis provides a mixing among the fermions which is the manifestation of the weak interactions. The mixing among the weak eigenstates of the down-type quarks (d' , s' , b') and the corresponding mass eigenstates d , s , b is characterized by the known Cabibbo-Kobayashi-Maskawa (CKM) matrix [24]. The off-diagonal elements of the CKM matrix explain that W bosons can couple to two quarks belonging to two different generations. The CKM matrix has four parameters; three mixing angles that control the mixing among each generation pair and one complex phase responsible for CP-violating⁵ phenomena.

Introducing the expansion described in Equation 2.17 to the Yukawa Lagrangian in

5. The Charge Parity (CP) symmetry is a combination of the charge conjugation symmetry and the parity symmetry which states that the laws of physics should be the same if a particle or a system of particles are interchanged with respective antiparticles (C symmetry) and when its spatial coordinates are inverted (P symmetry).

Equation 2.28, predicts the tree level mass of the fermions to be:

$$m_f = y_f \frac{v}{\sqrt{2}}, \quad (2.27)$$

where f denotes the fermions of the theory.

Table 2.1 summarizes the intensity of the couplings of the Higgs bosons to the vector gauge bosons ($V \equiv W, Z$), to fermions (f), and to itself. Equation 2.27 shows that the particles masses are proportional to Higgs couplings. Therefore, the Higgs boson will be more favorably produced in association with heavy particles, and will decay more favorably into the heaviest particles that are kinematically allowed.

Coupling	Intensity
$Hf\bar{f}$	m_f/v
HVV	$2m_V^2/v$
$HHVV$	$2m_V^2/v^2$
HHH	$3m_H^2/v$
$HHHH$	$3m_H^2/v^2$

Table 2.1: The Higgs boson couplings to fermions (f), vector gauge bosons (W, Z), and the Higgs self-coupling in the SM.

2.4 The Higgs Boson

The Higgs boson according to the Standard Model, is a neutral particle with spin zero whose mass is a free parameter to be determined experimentally. Extensive amount work at the LHC, in both the ATLAS [1] and CMS [2] experiments, led to the discovery of a new particle with a mass around 125 GeV, which was announced on the 4th of July in 2012. All possible production and decay rates need to be measured and compared with the predictions of the SM, in order to determine the properties of this newly discovered particle. In the following, both the Higgs boson production mechanism and decay modes are explained.

2.4.1 Production Mechanisms of the Higgs Boson

At the LHC, there are four highest cross-section production mechanisms of a Higgs boson with a mass of 125 GeV. They are described here, in decreasing order of production cross section.

The dominant production mechanism is via the gluon fusion ($gg \rightarrow H$ or simply ggH) process, where two merging gluons create a quark loop resulting in the creation of a Higgs boson. This is the dominant production mode, with a cross section of about 43.92 pb at 13 TeV, due to the overwhelming presence of gluons in pp collisions. The leading order diagram for this process is shown in Figure 2.3 (a). This production is mainly mediated by virtual top- or bottom-quark loops because the matrix element is proportional to the squared Yukawa coupling ($y_q^2 \propto m_q^2$), and other lighter quarks loops are highly suppressed.

The second leading production process is the vector boson fusion (VBF) that occurs about one order of magnitude less often than gluon fusion. In this process, vector bosons V (W^\pm or Z^0) which are mediated from two scattering quarks, merge and create a Higgs boson. The presence of diagrams with a vertex connecting the bosons to the Higgs boson without being in a loop, as shown in Figure 2.3 (b), is referred to as direct coupling. In VBF the incoming quarks undergo a large momentum transfer, resulting in energetic jets in the forwards direction, allowing a direct measurement of the coupling of the Higgs bosons to vector bosons.

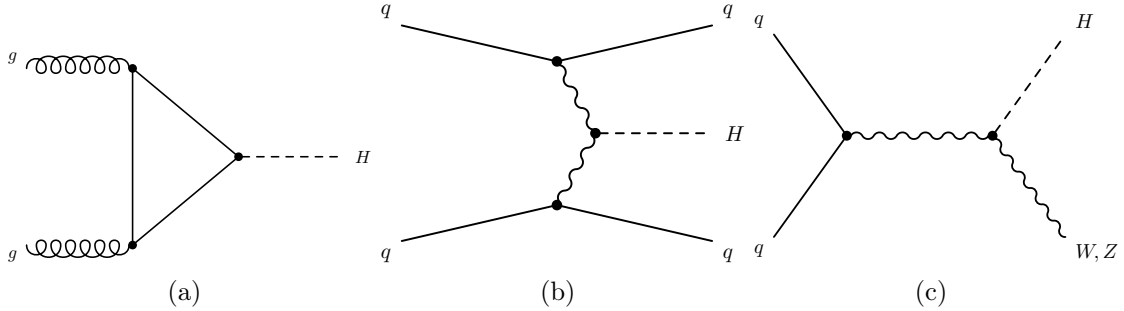


Figure 2.3: Examples of leading order Feynman diagrams for Higgs boson production via (a) gluon fusion, (b) vector boson fusion, and (c) Higgs boson in association with a vector boson production process.

The third production mode is the Higgs-strahlung, or associated production of the Higgs boson with vector bosons (VH). The Feynman diagrams for qq initiated process is

shown in Figure 2.3 (c). These production modes, being dominated by qq , allow to study the $H \rightarrow bb$ process since the leptonic decays of the additional vector bosons help reducing the multi-jet background.

The Higgs boson production in association with top-quarks ($t\bar{t}H$ or tH) is the rarest considered Higgs boson production mode here, which is suppressed by two orders of magnitude compared to gluon fusion. Figure 2.4 shows few Feynman diagrams for $t\bar{t}H$ that involve direct coupling of the Higgs boson to the top-quarks. The $t\bar{t}H$ production is the preferred process for the measurement of the top Higgs Yukawa coupling since it has a higher cross-section compared to the tH process.

The production of the Higgs boson in the (tH) process is mainly radiated from the top-quark, but it can also be radiated from the W boson propagator, causing an interference among these two diagram types. Therefore, the tH production rate is sensitive to the sign of the top Higgs Yukawa coupling. The sign of the Yukawa coupling is predicted to be positive in the SM producing a destructive interference. On the other hand, the sign can be negative in theories Beyond the Standard Model (BSM), resulting in constructive interference that could significantly enhance the production cross section.

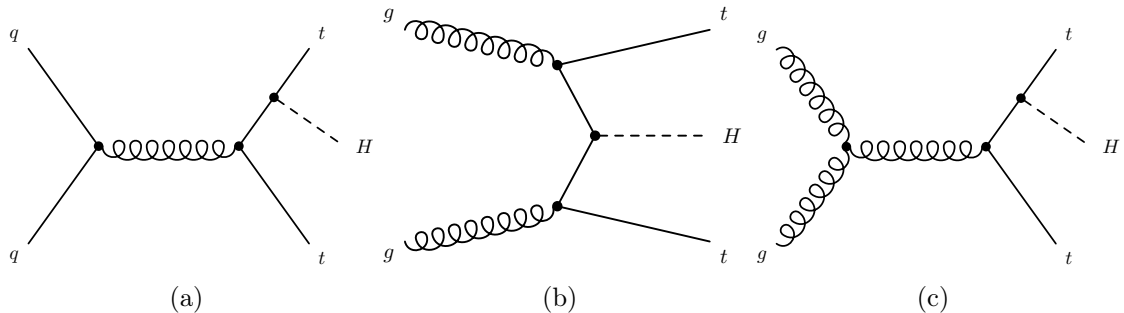


Figure 2.4: Examples of leading order Feynman diagrams for the production of a Higgs boson in association with top-quarks ($t\bar{t}H$).

A summary of the cross sections of the various production mechanism of a Higgs boson with a mass of 125 GeV as a function of the center-of-mass energy \sqrt{s} is shown in Figure 2.5. The $t\bar{t}H$ cross section at 13 TeV, is about 0.507 pb, increased by a factor of four compared to 8 TeV.

Analogous to $t\bar{t}H$ production, the Higgs boson can be also produced in association with bottom quarks. This process has a cross section surprisingly higher than that of $t\bar{t}H$

at lower center-of-mass energies, as shown in Figure 2.5.

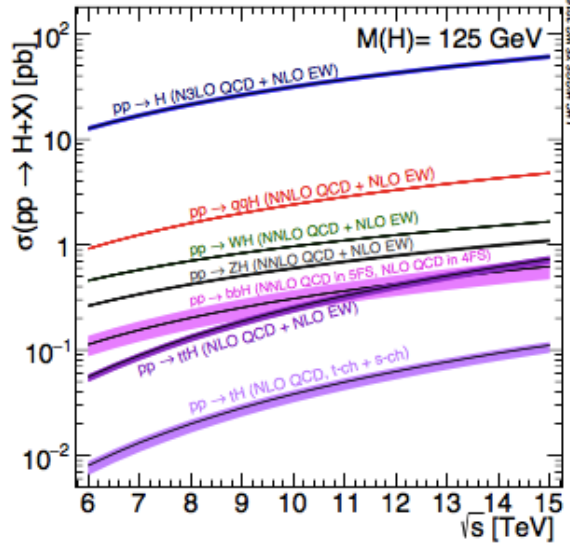


Figure 2.5: The SM Higgs boson production cross sections as a function of the center-of-mass energy \sqrt{s} . Note that the tH production cross section accounts for t -channel and s -channel only [18].

2.4.2 Higgs Boson Decays

The Higgs boson has a lifetime of some 10^{-22} s and is therefore only indirectly observed from its decay products. The decay widths into massive gauge bosons ($V = W, Z$) or fermions are proportional to the g_{HVV} and $g_{Hf\bar{f}}$ couplings respectively. Figure 2.6 shows the branching ratios of Higgs decay modes as a function of the mass of the Higgs boson and Table 2.2 shows the decay branching ratios of a 125 GeV Higgs boson. The Higgs boson will decay mostly to heavy particles such as pairs of electroweak gauge boson (W^\pm, Z) and into pairs of quarks and leptons (b, τ) as shown in Figure 2.7. The $H \rightarrow b\bar{b}$ channel has the largest branching ratio for $m_H = 125$ GeV.

As mentioned before, the Higgs boson does not couple to massless particles. Therefore, the decay modes in two photons or two gluons are induced through heavy particle loops as shown in Figure 2.8.

Distinguishing the Higgs boson signal from other processes with similar experimental signatures, referred to as background, is the main challenge in performing Higgs boson

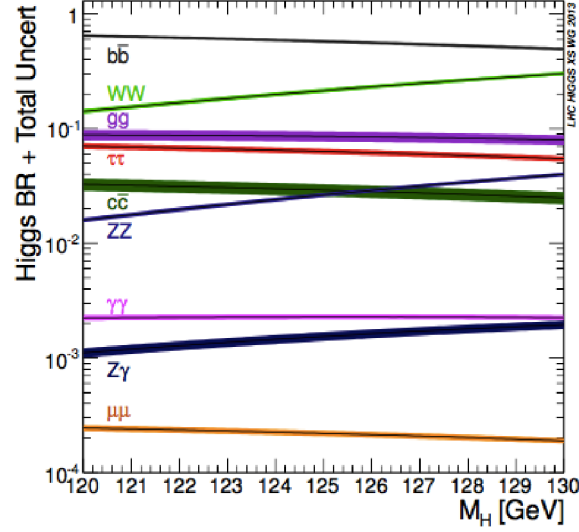


Figure 2.6: The branching ratios and their total uncertainty for the different SM Higgs boson decay modes for two different mass ranges from 120 up to 130 GeV [18].

Decay channel	Branching ratio [%]
$H \rightarrow bb$	58.2
$H \rightarrow WW$	21.4
$H \rightarrow gg$	8.19
$H \rightarrow \tau\tau$	6.27
$H \rightarrow cc$	2.89
$H \rightarrow ZZ$	2.62
$H \rightarrow \gamma\gamma$	0.227
$H \rightarrow Z\gamma$	0.153
$H \rightarrow \mu\mu$	0.022

Table 2.2: Branching ratios of the Higgs boson with a mass of 125 GeV [18].

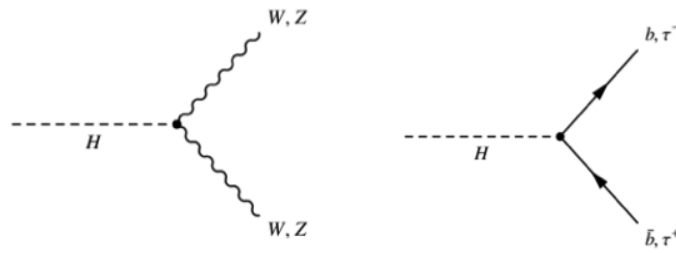


Figure 2.7: The Higgs boson decays to W , Z bosons and to fermions.

measurements. Different Higgs boson decay modes have different background compositions

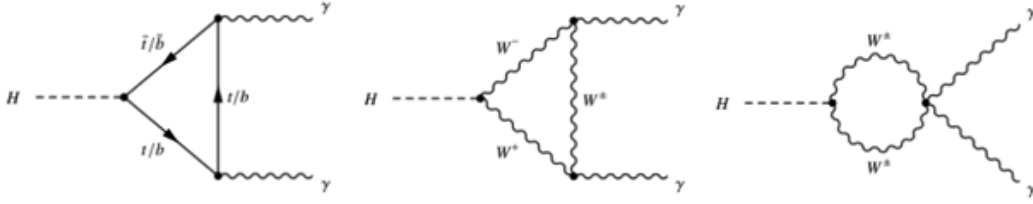


Figure 2.8: The Higgs decay to $\gamma\gamma$ mediated by heavy quark loops and W boson.

and experimental challenges.

2.4.3 Properties of the Higgs Boson

The Higgs boson was discovered in 2012 using the ZZ , $\gamma\gamma$ decay channels [1, 2, 25]. The combined measurement of the Higgs mass from the ATLAS and CMS collaborations with the full Run 1 dataset was found to be

$$m_H = 125.09 \pm 0.21(\text{stat.}) \pm 0.11(\text{syst.}) \text{ GeV}. \quad (2.28)$$

The first observation of fermionic decays was later seen in $H \rightarrow \tau\tau$ decays [26, 27], and evidence for the Higgs decay to bb was found in 2017 through the VH production mode [28, 29].

In the SM, the Higgs boson is a spin-0 and CP-even particle ($J^P = 0^+$). This is tested against several alternative spin-parity hypothesis based on the kinematic properties of the $H \rightarrow \gamma\gamma$, $H \rightarrow ZZ^*$, and $H \rightarrow WW^* \rightarrow l\nu l\nu$, which differ depending on J^P . The spin 1 and 2 hypothesis are rejected at respective confidence levels higher than 99.7% and 99.9%, using the 8 TeV ATLAS data [30]. Similar studies were performed in the CMS collaboration [31]. These preliminary results show evidence of the spin 0 nature of the Higgs boson, which is compatible with the SM, and also show a preference for the even parity predicted by the SM.

2.5 The Top-Quark

The top-quark was discovered in 1995 by the CDF [32] and D0 [33] collaborations. It belongs to the third generation of quarks, together with the bottom quark. The top-quark

is the heaviest particle in the SM with a mass of 173.2 ± 0.9 GeV [20]. Moreover, the top-quark has other unique and special properties, such as the large value of its width (1.41 ± 0.17 GeV [20]) that causes it to have a very short lifetime of about 5.0×10^{-25} s. This implies that the top-quark decays before any hadronization effect can take place. This allows us to directly detect spin information transferred to its decay products undiluted by non-perturbative effects.

An important consequence of its high mass is the strong Yukawa coupling to the Higgs boson, which is very close to 1 according to Equation 2.27:

$$y_t = \sqrt{2} \frac{m_t}{v} \approx 1. \quad (2.29)$$

This might be a coincidence but could also have a deeper reason, and any experimental deviations could hint for new physics beyond the SM. In the following, the $t\bar{t}$ production and decay modes will be described since they are a main ingredient of the analysis presented in this thesis.

2.5.1 Top-Quark Pair Production

Top pair ($t\bar{t}$) production is the most common mode to produce top-quarks at the LHC, which is done via $q\bar{q}$ annihilation or gluon gluon fusion. Figure 2.9 shows the four leading-order (LO) Feynman diagrams of top-quark pair production, in which one is produced through $q\bar{q}$ annihilation and three through gluon fusion.

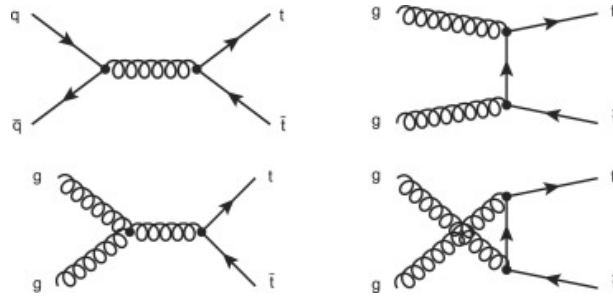


Figure 2.9: Four LO Feynman diagrams of the top-quark pair production through the strong interaction.

The dominant production mechanism at the Tevatron $p\bar{p}$ collider was the $q\bar{q}$ annihilation ($\approx 85\%$ of $t\bar{t}$ cross section) in which the collisions happen mainly between the valence quarks

from the proton and the anti-proton. While at the LHC, 80 – 90% of the $t\bar{t}$ pairs are produced via gluon fusion, depending on the center-of-mass energy.

Figure 2.10 shows both the theoretical and the measurements of the $t\bar{t}$ production cross section as function of the center-of-mass energies. The theoretical computation is made at next-to-next-to-leading order (NNLO) in α_s and with next-to-next-to-leading logarithm (NNLL) soft-gluon resummation [34–39]. Details on α_s and resummation will be discussed in Chapter 3. For a top-quark mass of $m_t = 173.2$ GeV, the cross section is $\sigma_{t\bar{t}}(8 \text{ TeV}) = 247.7^{+13.1}_{-14.3}$ pb, and $\sigma_{t\bar{t}}(13 \text{ TeV}) = 816.0^{+39.5}_{-44.7}$ pb [40] in which the uncertainty comes from variations of the renormalization and factorization scales as well as uncertainty associated with the parton distribution functions.

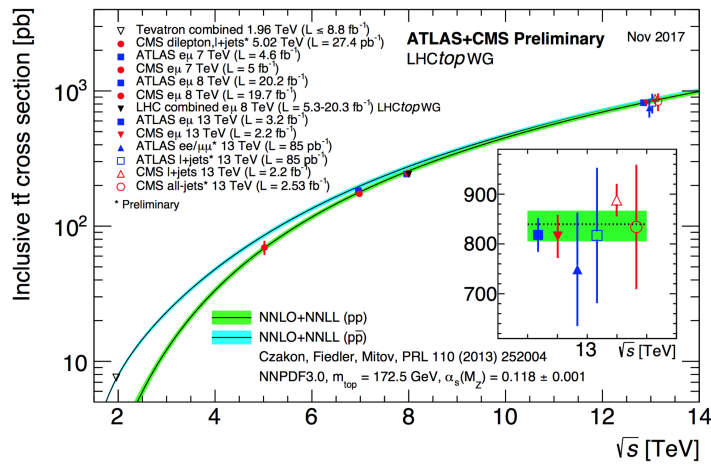


Figure 2.10: Summary of the LHC and Tevatron measurements of the top-pair production cross-section as a function of the center-of-mass energy compared to the NNLO QCD calculation complemented with NNLL resummation (top++2.0). The theory band reflects the uncertainties arising from the renormalization and factorization scale, parton density functions and the strong coupling. The theoretical calculations and the measurements are quoted at $m_t = 172.5$ GeV [40].

2.5.2 Decay of the Top-Quark

The top-quark decays almost exclusively into a b -quark and a W boson. Furthermore, the b -quark hadronizes and the W boson decays either hadronically into a pair of light quarks $q\bar{q}$ ($u\bar{d}$ or $c\bar{s}$, 68%) or leptonically into a charged lepton and the corresponding

neutrino ($W \rightarrow l\nu_l$, 32%). The final $t\bar{t}$ state is determined upon the number and flavor of the decay products of the two W bosons present in the event. The different $t\bar{t}$ signatures are displayed in Figure 2.11 and listed as:

- **fully hadronic:** refers to the decay of $t\bar{t} \rightarrow b\bar{b} W^+W^- \rightarrow b\bar{b} q\bar{q}' q''\bar{q}'''$ which corresponds to $\approx 46\%$ of the branching ratio. In the LO picture, six jets are expected two of which are b-jets.
- **dilepton:** refers to the decay $t\bar{t} \rightarrow b\bar{b} W^+W^- \rightarrow b\bar{b} l\nu_l l'\nu_{l'}$ which corresponds to $\approx 9\%$ of the branching ratio. In the LO picture two b-jets, two opposite sign charged leptons and two neutrinos resulting in large missing transverse energy are expected.
- **single lepton:** refers to the decay $t\bar{t} \rightarrow b\bar{b} W^+W^- \rightarrow b\bar{b} q\bar{q}' l\nu_l$ which corresponds to $\approx 45\%$ of the branching ratio. In the LO picture, four jets are expected two of which are b-jets, one charged electron or muon and one neutrino resulting in missing transverse energy.

Note that the τ lepton decays leptonically or hadronically and it is usually treated separately. The leptonic decay of τ leptons result in the same signatures as described above and are experimentally included into the dilepton and single lepton channels.

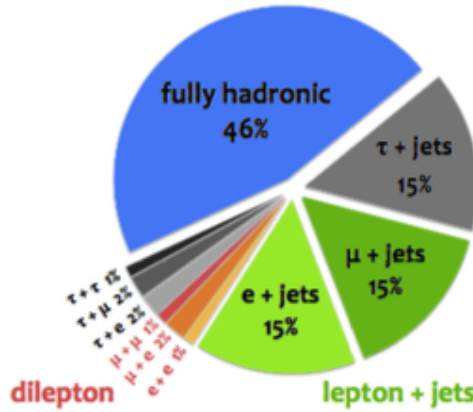


Figure 2.11: Pie chart showing the branching ratios (BR) of a top-antitop quark pair. The blue color represents the fully hadronic BR of 46% (56% when including hadronic decaying τ), different shades of red represent the dileptonic BR without τ lepton with a total of 4% (6.4% when including leptonic decaying τ), and different shades of green show the lepton (e or μ) + jets BR of 30% (36% when including leptonic decaying τ).

2.6 Direct Measurement of the Top Higgs Yukawa Coupling

Determining the Yukawa coupling between the Higgs boson and the top-quark y_t is possible by measuring the cross section of the gluon fusion production process and the $H \rightarrow \gamma\gamma$ decay mode. In these processes a sizeable contribution arises from a top-quark loop, as shown in Figure 2.12 (a) and (b), and assumes that no new physics contributes with additional induced loops in the measurement of y_t . However, the only experimentally accessible process now is the one in which y_t can be accessed directly through the production of a top-quark pair in association with a Higgs boson ($t\bar{t}H$), as shown in Figure 2.12 (c), and it is the subject of this thesis.

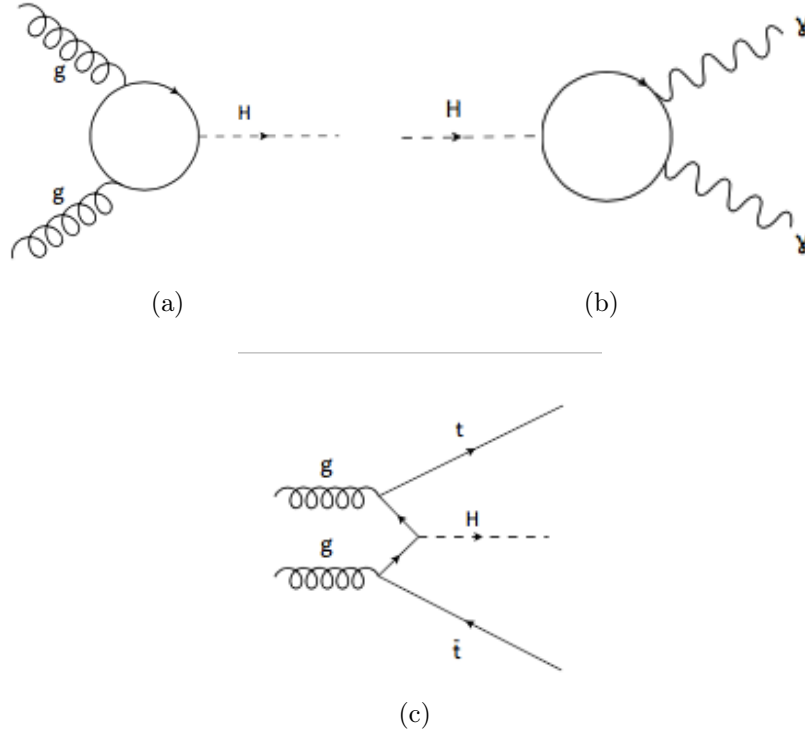


Figure 2.12: Feynman diagram for (a) the effective gluon fusion vertex, (b) the effective photon vertex ($\gamma\gamma H$), and (c) the production of a top-quark pair in association with a Higgs boson ($t\bar{t}H$).

The search of the production of the Higgs boson in association with a pair of top-quark is expressed in terms of the signal strength parameter $\mu_{t\bar{t}H}$, which is defined as the ratio of the observed to the expected number of signal events assuming a SM cross section.

Both ATLAS and CMS collaborations have searched for the production of $t\bar{t}H$ in pp collisions at the LHC using the data collected at a center-of-mass energy of 7, 8, and 13 TeV in the $H \rightarrow WW^*, \tau\tau, b\bar{b}$ and $\gamma\gamma$ decays. The combination of ATLAS and CMS, using the 7 and 8 TeV data, results yields a best fit of $\mu_{t\bar{t}H} = \sigma/\sigma_{SM} = 2.3^{+0.7}_{-0.6}$, with an excess over the SM expectation ($\mu_{t\bar{t}H} = 1$) mainly driven by the multilepton final states [25].

Chapter 3

SIMULATION OF PARTICLE INTERACTIONS

To measure the $t\bar{t}H(H \rightarrow b\bar{b})$ process, it is important to estimate the detector acceptance and to optimize the sensitivity using simulated events. Simulated data is derived through the Monte Carlo (MC) method in which repeated random sampling of variables from probability distributions based on phase-space integrations of matrix element calculations, are used to model the signal and background processes and to provide theoretical uncertainties using the most up to date theoretical knowledge. MC events are simulated over three primary steps: matrix element simulation of the hard interaction, parton showering and hadronization, and simulation of the detector response.

The simulation of pp collisions requires a detailed description of physics processes including a wide range of energy scales. At the high-energy scale are deep-inelastic interactions between partons, calculated in perturbative Quantum Chromodynamics (QCD). Low energies include the evolution of partons into stable hadrons, which cannot be calculated perturbatively. Therefore, a key aspect in the simulation of pp collisions is the ability to factorize the different energy scales involved in the process. QCD is briefly discussed in Section 3.1, and the factorization theorem is explained in Section 3.2.

The steps involved in the simulation of a pp collisions in MC event generation are illustrated in Figure 3.1. The simulation starts with the calculation of the matrix element for the process of interest, known as the hard scattering process (e.g. $pp \rightarrow t\bar{t}H$), explained in Section 3.3. This is followed by mostly soft and collinear parton branchings which are simulated by the initial and final state parton shower algorithms, as explained in Section 3.4. The evolution of the parton shower is followed by the hadronization process, described in Section 3.5, which produces collimated bunches of hadrons including their decay products. They represent the final collection of energy deposits in the detector referred to as jets. Secondary interactions forming the underlying event are discussed in Section 3.6. A general overview of the MC generators used in the $t\bar{t}H(H \rightarrow b\bar{b})$ search is presented in Section 3.7, and the ATLAS detector simulation is briefly discussed in Section 3.8.

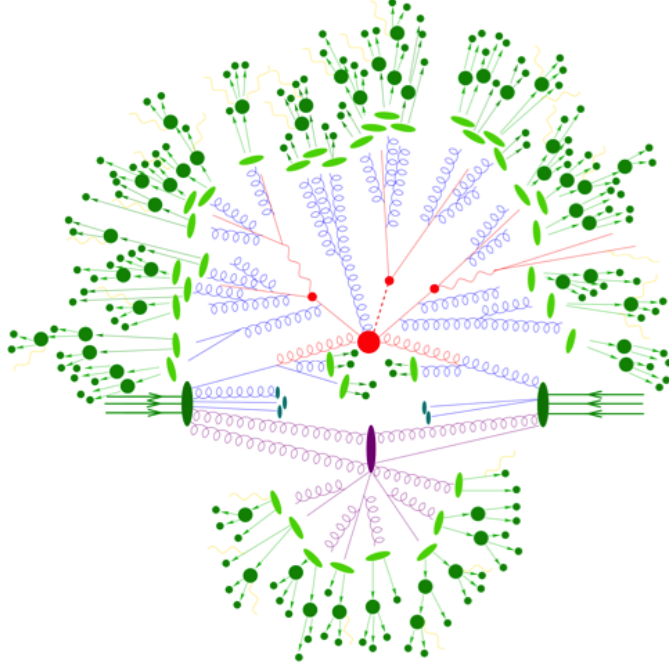


Figure 3.1: Representation of a $t\bar{t}H$ production event containing all steps in the event generation chain. The two incoming protons with three partons assuming no transfer momentum of partons, are represented by the two green ovals. The red circle in the center represents the large momentum transfers in the primary hard scattering process, surrounded by a tree-like structure describing Bremsstrahlung as simulated by parton shower (initial and final state parton showers are denoted in blue and red respectively). The purple oval indicates a secondary interaction between other partons of the proton involving smaller momentum transfers. Light green blobs represent the parton-to-hadron transitions, the dark green blobs describe hadron decays, and the yellow lines indicate soft photon radiation [41].

3.1 Quantum Chromodynamics

QCD, briefly mentioned in Chapter 2, is a quantum field theory that describes the strong interaction of color charged particles such as quarks and gluons. Two important aspects underly the modeling of QCD known as *confinement* and *asymptotic freedom*. Confinement specifies that quarks do not exist in isolation but form colorless compound hadrons [42]. Asymptotic freedom describes the phenomenon that the strong force becomes asymptotically weaker at smaller distances or higher energies to the point that two partons are hardly interacting. Both aspects are associated to the running of the strong coupling constant. Figure 3.2 shows the strong coupling constant of QCD α_s and its dependence

on the energy scale Q of the interaction. It is shown that the strength of the strong force changes with the energy scale Q . The coupling is large for small scales, corresponding to large distances in which the theory is non-perturbative. For example, α_s is large for energies of the order of the proton mass. Therefore, the principle of confinement dominates in the low energy regions [42], since the large size of α_s is responsible for that. On the other hand, for high scales that correspond to small distances, α_s is small and the process can be calculated perturbatively (asymptotic freedom) [43].

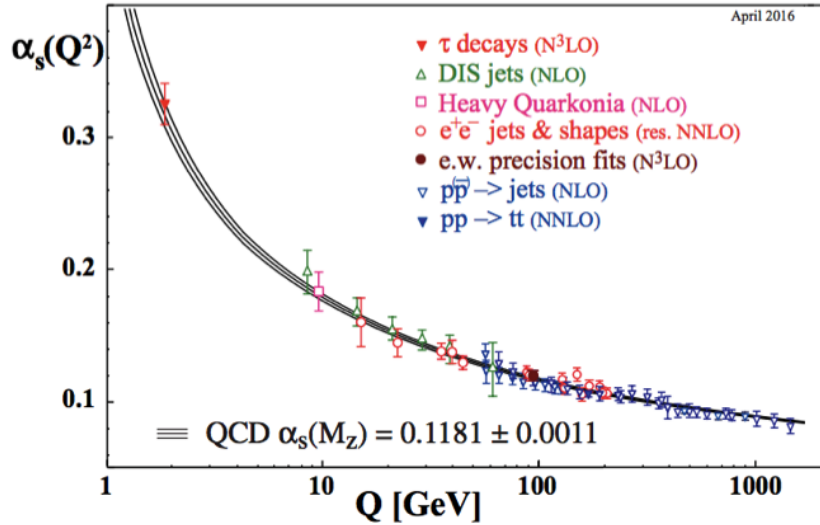


Figure 3.2: The QCD running coupling constant α_s as a function of the energy scale Q [20]. The shape of the running is predicted by the SU(3) theory, while the level is determined by experiment for fixed values of Q . The world average of α_s measured at the energy scale equal to the mass of the Z boson is illustrated. The respective degrees of QCD perturbation theory used in the extraction of α_s is indicated in brackets (NLO: next-to-leading order, NNLO: next-to-next-to leading order, res. NNLO: NNLO matched with resummed next-to-leading logs, and N³LO: next-to-NNLO).

Self-interaction in QCD theory that could lead to ultraviolet (UV) divergences are compensated for by using *regularization schemes*. This procedure is known as *renormalization* [44] that results in UV-finite cross sections with additional parameters that have to be introduced when applying the renormalization conditions. An arbitrary renormalization scale μ_R and a factorization scale μ_F are introduced to scale the finite set of parameters in the QCD theory to counteract divergent contributions. The value of μ_R is usually chosen to be equal to the factorization scale μ_F , which defines the separation between

the perturbative treatment of the short-distance interactions and the modeling of the long-distance interactions by means of the parton distribution functions.

3.2 The Factorization Theorem: PDFs and the DGLAP Equations

Interactions among the components of the proton, called partons, occur during pp collisions at the LHC. The partons are either one of the three valence-quarks (uud) or spontaneously produced non-valence quark-anti-quark pairs and gluons that emerge from the strong interaction between the valence quarks. The partons behave as asymptotically free particles at high energy, where a perturbative description is applied. The inclusive cross section for a process such as $pp \rightarrow X$, illustrated in Figure 3.3, is defined in terms of the cross section $\hat{\sigma}_{ab \rightarrow X}$ for the partonic processes according to the factorization theorem [45] as:

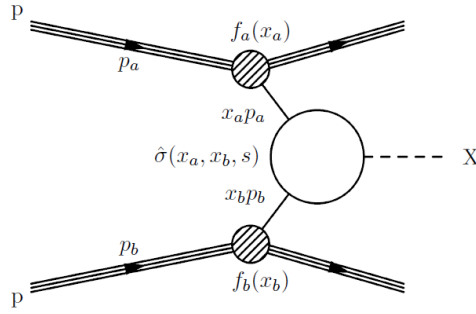


Figure 3.3: Illustration of a generic hard scattering process. The partons, obtained from the colliding pp pair, carry a momentum fraction (x_a, x_b) with respect to the proton energy $(p_a, p_b, \text{ respectively})$ described by a parton distribution function. The scattering of the partons is computed perturbatively. Therefore, the kinematic properties of the final state object X are predicted.

$$\sigma_{pp \rightarrow X} = \sum_{a,b} \int dx_a dx_b f_a(x_a, \mu_F^2) f_b(x_b, \mu_F^2) \hat{\sigma}_{ab \rightarrow X}(x_a p_a, x_b p_b, \mu_R^2, \mu_F^2), \quad (3.1)$$

where the sum runs over the partons types (a, b) that can initiate the process. The parton density function (PDF), $f_i(x_i, \mu_F^2)$, describes the momentum distribution of partons within

a proton [46]. A PDF gives the probability density of finding a parton of type (i) within the proton, carrying a fraction of the proton's momentum (x_i) for an energy scale (Q). PDFs cannot be predicted directly due to the non-perturbative QCD description of the strong interaction among partons inside a hadron. Instead, they are calculated using measurements from several hadron colliders and deep inelastic scattering experiments such as H1 and ZEUS at the electron-proton HERA [47] collider.

The measurements are only feasible for certain Q^2 scales and must be extrapolated to the regime of interest. The energy dependence of the PDFs is described by the DGLAP evolution equation [48–50] as:

$$\frac{\partial q_i(x, Q^2)}{\partial \log Q^2} = \frac{\alpha_s}{2\pi} \int_x^1 \frac{dz}{z} \{ P_{q_i q_j}(z, \alpha_s) q_j(\frac{x}{z}, Q^2) + P_{q_i g}(z, \alpha_s) g(\frac{x}{z}, Q^2) \}. \quad (3.2)$$

$$\frac{\partial g(x, Q^2)}{\partial \log Q^2} = \frac{\alpha_s}{2\pi} \int_x^1 \frac{dz}{z} \{ P_{g q_j}(z, \alpha_s) q_j(\frac{x}{z}, Q^2) + P_{g g}(z, \alpha_s) g(\frac{x}{z}, Q^2) \}, \quad (3.3)$$

where $q_i(x, Q^2)$ in Equation 3.2 is the quark PDF, $g(x, Q^2)$ in Equation 3.3 is the gluon PDF, and $P_{ab}(z, Q^2)$ are the splitting functions that can be expanded in powers of the running coupling [51]. There are no equations for the evolution in x , which is obtained from fits to experimental data. Various collaborations constantly work to improve the PDF fits using the most recent data. The following PDF sets are commonly used at the LHC and in the analysis presented in this thesis: CTEQ [52], NNPDF [53], and MSTW [54]. Since the PDF groups use slightly different assumptions for the DGLAP equation, different groups are used to estimate the theoretical uncertainty. An example of the NNPDF2.3 PDF set is shown in Figure 3.4. It includes previous LHC among many other datasets and is calculated at NNLO accuracy, taking into account terms up to order α_s^3 in the DGLAP equations. It is worth mentioning that the scale $Q^2 = 10^4$ GeV in Figure 3.4 (b) corresponds to the typical momentum transfer for Higgs boson production and that at 13 TeV the most likely values of x of the incoming partons are around 10^{-2} (assuming symmetric collisions), so Higgs boson is dominantly produced by gluon gluon fusion and the valence quarks play a very minor role. Generally, the order of the PDF calculation should be equivalent to the order of the matrix elements used in the hard process part of the MC calculation.

According to the QCD factorization theorem [45], the hard scattering can be considered

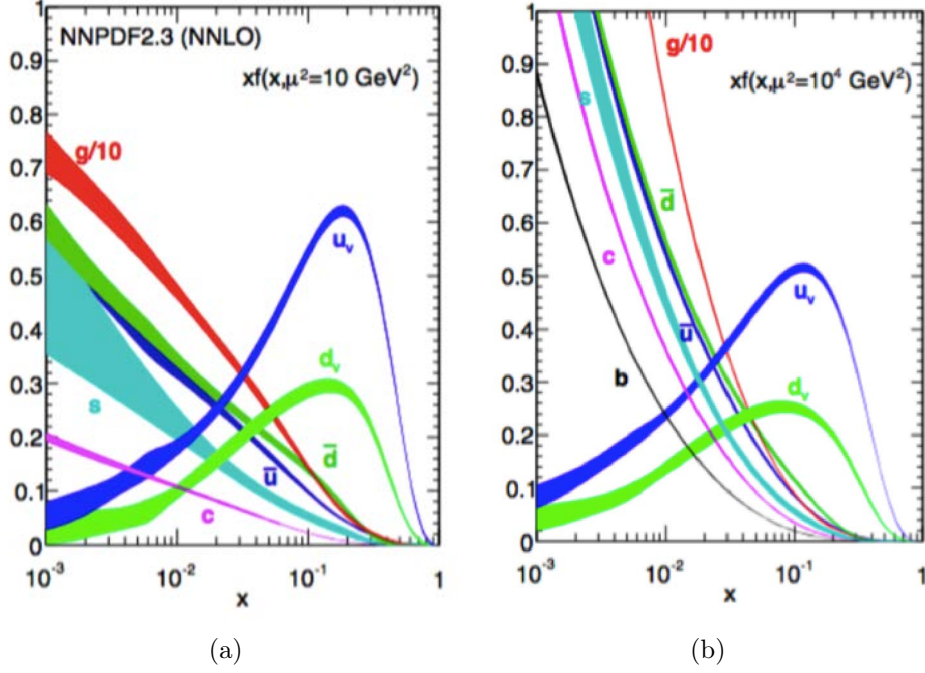


Figure 3.4: Example of NNLO parton distribution functions for various parton flavors as a function of fractional parton momentum x at a scale of (a) 10 GeV^2 and (b) 10^4 GeV^2 . Results from NNPDF2.3 set uses also data from the LHC [55, 56].

separately from the PDFs. The factorization scale μ_F , denoted by Q in Equations 3.2 and 3.3, defines the boundary between the kinematic region where emissions are treated as part of the hard scatter and the region where emissions are included in the PDF.

3.3 Matrix Element

The simulation of particle interactions depends on the calculation of the matrix element that describes the transition of an initial to a final state. The matrix element describes the primary scattering of the partons, known as the hard process, at the highest momentum scales. The total inclusive cross section for producing any final state (X) from a hadron collision is expressed to all orders in perturbation theory as:

$$\sigma = \sum_{k=0}^{\infty} \int_{m+k} d\Phi_{m+k} \left| \sum_{l=0}^{\infty} \mathcal{M}_{m+k}^{(l)}(\Phi_{m+k}) \right|^2 \quad (3.4)$$

with

$$d\Phi_{m+k} = SF \prod_{f=0}^{m+k} \frac{d^3\vec{p}_f}{(2\pi)^3} \frac{1}{2E_f}, \quad (3.5)$$

where m is the number of particles in final-state X , k denotes the number of additional real emissions, l is the number of virtual correction loops, S describes the symmetry factor that appears for groups of identical final-state particles, F represents the flux factor, and $\mathcal{M}_{m+k}^{(l)}$ is the scattering amplitude corresponding to the sum of the Feynman diagrams with l loops and $m+k$ final-state particles.

Figure 3.5 shows an example of three Feynman diagrams for a $t\bar{t}$ final state at tree level ($k=0, l=0$), first emission ($k=1, l=0$), and including a virtual correction ($k=0, l=1$).

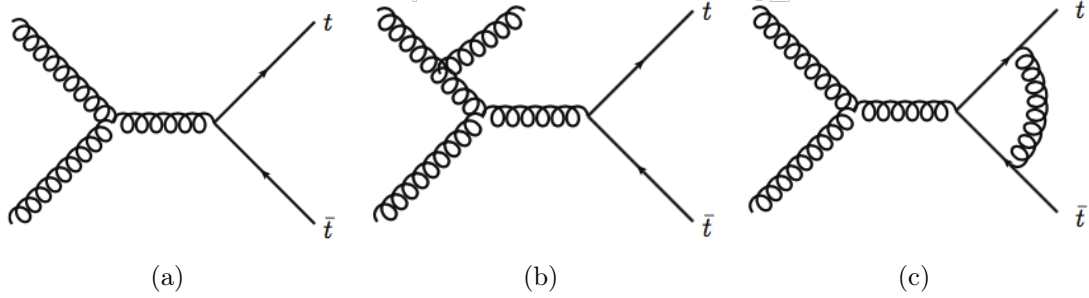


Figure 3.5: Example of Feynman diagrams of a $t\bar{t}$ production (a) at leading order, (b) for a first real emission, and (c) for a first virtual correction.

3.4 Parton Shower

The parton shower (PS) refers to partons above the PS cut-off ($\simeq 1-2$) GeV that undergo QCD radiation of gluons and photons, as indicated by the red lines in Figure 3.1. The parton shower approximates the contributions of higher order in perturbation theory, to mimic a complete final state. MC generators which model the parton shower, simulate the successive emission of gluons and quarks from the partons in the initial and final state. These simulations consider independent parton emissions and do not include virtual corrections, which makes them approximate. The parton shower contribution to the hard process cross section is predicted by only considering the dominant contribution to each

order. Starting with a differential cross section for n particles ($d\sigma_n$), a differential cross section for $n + 1$ particles is calculated according to the following equation:

$$d\sigma_{n+1} \approx d\sigma_n dP_i(z, q^2) \approx d\sigma_n \frac{\alpha_s}{2\pi} \frac{dq^2}{q^2} dz P_{ji}(z). \quad (3.6)$$

where $dP_i(z, q^2)$ is the probability that parton i will split into two partons with parton j that carries a fraction z of the momentum of parton i , and q^2 denotes the evolution variable of the parton shower. Note that the evolution variables differ between different parton shower generators, such as the invariant mass or the transverse mass.

Figure 3.6 illustrates the above mentioned process. Three possible processes for QCD emission or splitting can occur: $g \rightarrow gq$, $g \rightarrow gg$, $g \rightarrow q\bar{q}$. The simulation algorithm develops the shower by applying Equation 3.6 iteratively, to evolve the event from the large scale associated to the hard scattering to the lower parton shower cut-off scale, where perturbation theory breaks down and phenomenological hadronization models take over.

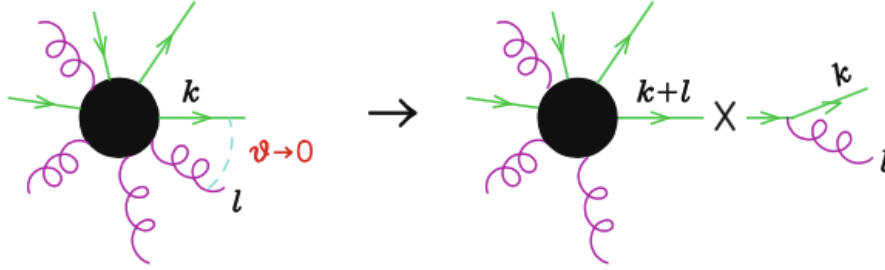


Figure 3.6: A representation of a splitting of an n -parton process to an $n+1$ -parton process. This figure is taken from [57]

The parton shower is implemented in the MC algorithm via the so-called Sudakov form factors:

$$\Delta_i(q_1^2, q_2^2) = \exp\left(-\sum_i \int_{q_2^2}^{q_1^2} \int_{z_{\min}}^{z_{\max}} dP_i(z, q^2)\right). \quad (3.7)$$

The Sudakov form factors of a parton (i), Δ_i , represented in Equation 3.7 describe the probability that a parton evolves from an initial scale q_1 to a lower scale q_2 without splitting. The algorithm implemented in MC simulations operates in the following steps:

- Given the initial energy scale Q^2 , partons emit radiation at a scale q_2 determined by

Equation 3.7.

- If the value of q_2^2 is below the hadronization scale, $q_2^2 < Q_0^2 \approx 1 \text{ GeV}^2$, no further emission occurs. Therefore, the shower development is terminated and hadronization takes place.
- Otherwise, the procedure is repeated where further emissions occur for each new parton using q_2^2 as the initial scale.

Initial-state showers occur when the radiation is emitted by the incoming partons before the scatter.

3.5 Hadronization

The parton shower is terminated once the generated partons have a virtuality below the PS cut-off scale. Perturbation theory becomes invalid at low energies and large distances because of the increasingly large coupling α_s , which leads to quark confinement. Individual partons start to hadronize into colorless baryons and mesons. Hadrons are build out of partons that bind together. These hadrons might be excited and also decay into many lower-energy states. Hadronization is typically simulated through either the cluster model [58,59] or the Lund string model [60,61].

The cluster model relies on the concept of preconfinement [62], which starts with the non-perturbative splitting of gluons into color-singlet $q\bar{q}$ pairs, as illustrated in Figure 3.7 (a). Color-singlet combinations are then grouped into clusters which are individually evaluated to predict daughter hadrons depending on the density states and quantum properties. The heaviest clusters can decay and split into smaller clusters. Clusters with a mass below 3-4 GeV are transformed into hadrons through a two-body decay [63].

The Lund string model, as illustrated in Figure 3.7 (b), gives a more continuous description of the hadronization. It describes the color flux between the stretched $q\bar{q}$ pair, where the potential among the partons is proportional to their distance. The potential is thought of as a virtual color flux string. When the distance and consequently the potential is adequately large, the string breaks and forms a new $q\bar{q}$ pair. Radiated gluons are considered as kinks along the string, carrying momentum. The model needs extra

parameters to define the transverse momentum distribution of the hadrons and heavy particle suppression.

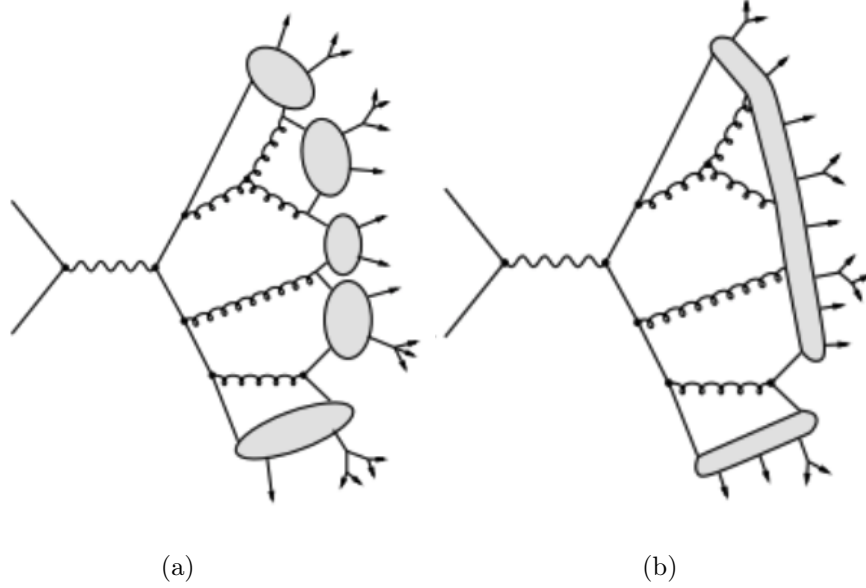


Figure 3.7: Sketches of (a) the cluster hadronization model where individual color-singlets are considered individually, and (b) the Lund string hadronization model.

3.6 Underlying Event

In addition to the hard process, secondary interactions between remnant partons in the incoming protons can happen, producing the underlying event (UE). The UE, indicated in the purple ovals in Figure 3.1, consists of the interactions between the remnant partons or the breakup of the beam remnants, i.e. the colored rest of the proton breakup, and the multiple parton interaction (MPI) that generates multiple distinct scatters. Due to the low energy scale of these processes, their modeling relies on phenomenological models with free parameters, which need to be tuned based on experimental data [64].

3.7 Monte Carlo Generators

The following gives an overview of the specific MC generators used for the studies presented in this thesis. Different MC generators implement various theoretical models and the

selection presented here is chosen in order to cover the theoretical uncertainty of the current understanding.

Matrix Element Generators

- POWHEG-BOX [65] is a NLO parton-level event generator computing matrix element in perturbative QCD using the POWHEG method [66] to match the matrix element calculation with the parton shower.
- MADGRAPH5_AMC@NLO [67] is a MC generator known for the automated computation of the matrix element at LO and NLO. The NLO calculation depends on the MC@NLO method [68] to match the matrix element calculation with the parton shower.

Multi-purpose Generators

The following multi-purpose generators have a specific implementation of the underlying event and beam remnants model. In the analysis presented in this thesis, they are only used for the parton shower, hadronization, and the underlying event modeling. They are interfaced with the matrix element generators.

- PYTHIA [69] is a multi-purpose MC generator using PS with emissions ordered in transverse momentum. It models the hadronization based on the Lund string model.
- HERWIG 7 [70, 71] is a multi-purpose MC generator. It uses PS with emissions ordered in opening angle that includes color-coherence effects with special description of radiation from heavy particle. It models the hadronization based on the cluster model.

EVTGEN [72], is a generator that runs after the parton shower and hadronization in the above mentioned PYTHIA or HERWIG 7 samples. EVTGEN has a detailed description of the physics of B -mesons, which includes detailed models for semileptonic decays, CP-violating decays, and produces accurate results for angular distributions in sequential decays, including all correlations.

Multi-purpose Generators with NLO Matrix Element

- SHERPA [73] is a NLO/LO multi-purpose MC generator used for many final states. It contains a parton shower algorithm based on the Catani-Seymour dipole formalism [74]. It can be interfaced with additional libraries to compute loop amplitudes. SHERPA interfaced with OPENLOOPS [75] is used to model the $t\bar{t} + b\bar{b}$ process at NLO which represents the largest background for the analysis presented in this thesis.

3.8 ATLAS Simulation

The result of the MC generator calculations is a list of four-vectors of all stable¹ particles produced in the event, after hadronization and decay of the intermediate unstable particles. These results are stored to study processes at the so-called stable particle level. In order to account for the detector response further simulation of the detector is needed. The detector simulation software, which models the interaction of the particles with the detector, is based on the GEANT4 framework [76].

The simulation of the interaction converts energy deposits into simulated electronic signals taking into account the geometry and response of the ATLAS detector. The detector simulation is highly CPU intensive. Out of all the different steps, the development of a particle shower in the calorimeter system of the ATLAS detector requires the largest amount of time to simulate. For example, the full detector simulation, referred to as Fullsim, for a single $t\bar{t}$ event, requires about 15 min of CPU time [77]. Therefore, a faster and less refined simulation, referred to as Atfast2 (AF2) [77], is developed to reduce the CPU time required to process the event by imposing a parametrized description of the particle showers in the calorimeters. The fully simulated samples generally provide higher precision and are favored for the main samples used in the analysis. However, AF2 samples are used in optimization studies or to assess theoretical systematic uncertainties.

In order to establish an accurate modeling of the detector effects including reconstruction and identification of physics objects, the simulated MC event samples are compared to data and corrected with multiplicative scale factors (SF), defined as:

$$\text{SF} = \frac{\epsilon_{\text{data}}}{\epsilon_{\text{MC}}}. \quad (3.8)$$

1. Stable refers to a final-state particle with mean lifetime $\tau = 3 \times 10^{-11}$ s.

where ϵ_{data} and ϵ_{MC} are measured in dedicated data calibration samples and in the equivalent MC simulation, respectively. Likewise, energy scale and resolution of the different physics objects in the simulated MC events are corrected to match the corresponding data measurements.

Chapter 4

THE LHC AND THE ATLAS DETECTOR

The Large Hadron Collider (LHC) [78] is a circular accelerator constructed to produce an extensive amount of TeV proton-proton (pp) and lead ions collisions which are measured by the ATLAS (A Toroidal LHC Apparatus) detector and the other experiments. Alongside ATLAS, there are six other experiments: CMS (Compact Muon Solenoid), LHCb (Large Hadron Collider beauty), ALICE (A Large Ion Collider Experiment), LHCf (LHC forward), TOTEM (TOTal Elastic and diffractive cross section Measurement) and MoEDAL (Monopole and Exotics Detector at the LHC). The ATLAS [79] and CMS [80] are general-purpose detectors designed to detect a wide range of signals and capable of studying Standard Model and beyond the Standard Model processes. ATLAS and CMS detectors are both designed to have an accurate electromagnetic calorimeter and high resolution tracking to identify and measure the four-momenta of all particle types, including leptons of all generations, photons, and jets. This thesis uses data collected by the ATLAS detector. Therefore, the following presents an overview of the LHC and the ATLAS detector.

4.1 The Large Hadron Collider

The LHC at the Conseil Européen pour la Recherche Nucléaire (CERN) is a chain of super conduction magnets constructed in the former tunnel of the Large Electron Positron (LEP) collider. Just after the dismantlement of the LEP in 2001 the construction of the ATLAS cavern and the other LHC experiments, CMS, ALICE, and LHCb started. The LHC tunnel has a circumference of approximately 27 km and lies between 45 m and 170 m under the French-Swiss borders at 1.4% inclination towards lac Léman in Geneva as demonstrated in Figure 4.1.

The LHC is a synchrotron designed to produce pp collisions at a center-of-mass energies of 14 TeV. In order to reach such high energy, hydrogen atoms are first ionized in an electric field and the resulting protons are sent through the linear accelerator LINAC2, where they are collected into bunches of roughly 1.15×10^{11} protons and accelerated up to 50 MeV. Then, the proton bunches are accelerated in sequence up to 1.4 GeV by the

CERN's Accelerator Complex

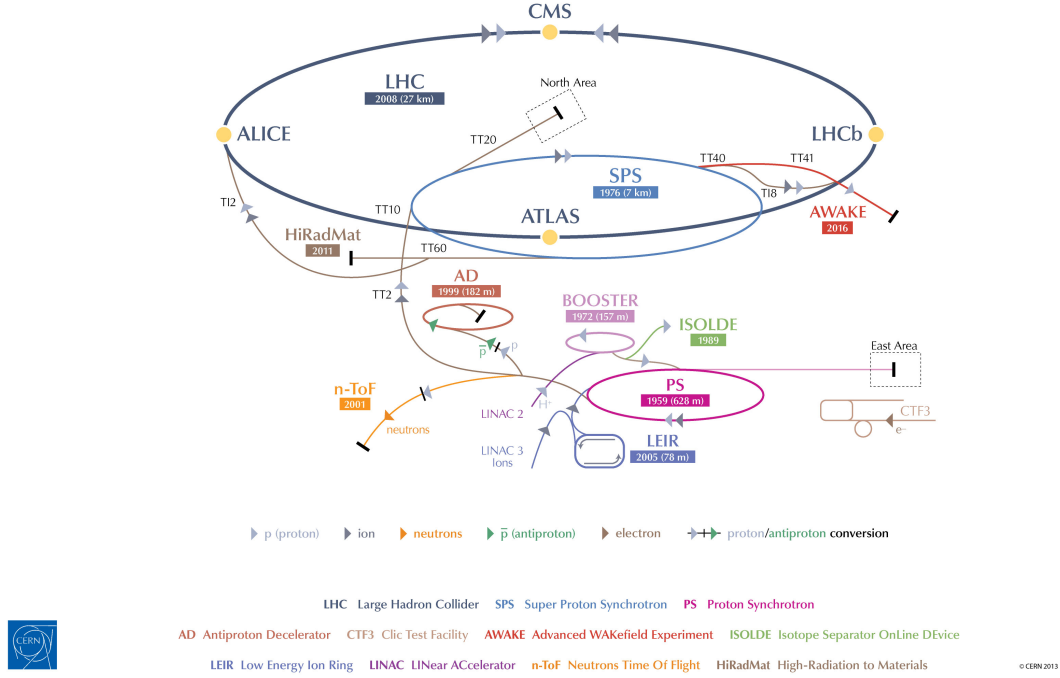


Figure 4.1: A cartoon of the CERN accelerator complex showing the LINAC2, BOOSTER, PS, SPS, and the LHC [81].

proton synchrotron (BOOSTER), up to 25 GeV by the proton Synchrotron (PS), and up to 450 GeV by the super proton synchrotron (SPS), after which they are injected in the LHC to be further accelerated to their final energies. It takes about 17 seconds to accelerate a single bunch from rest to 450 GeV, which is approximately 20 minutes to accelerate all bunches injected into the LHC beam. The proton bunches are further accelerated, squeezed into a condensed beam, and validated for physics. Collisions start once beams are declared stable and continue until the beam luminosity has decreased by roughly 50%, possibly up to 24 hours. A new fill starts once the bunches have lost a significant amount of their protons, impacting the data collection rate.

The planned operation of the LHC is divided into specific runs, each lasting several years and separated by long shutdowns for essential repairs and upgrades. Due to the quenching incident in 2008, the first LHC run, referred to as Run 1, was operated conservatively at half the design energy, collecting 5.08 fb^{-1} at 7 TeV during 2010 and 2011 and 21.3 fb^{-1}

at 8 TeV in 2012. The concept of luminosity, detailed in Section 4.2, is used to quantify the performance of a particle collider and the amount of pp collisions. Run 1 was a huge success which ended with the discovery of the Higgs boson. Following Run 1, the LHC had its first planned shutdown phase (LS1) that lasted until spring 2015. During this shutdown the LHC and its experiments went through several upgrades to prepare for operation at higher energies close to the design energy of 14 TeV with the purpose of collecting about 100 fb^{-1} of data. The second run of the LHC, referred to as Run 2, started in spring 2015 at a center-of-mass energy of 13 TeV. An integrated luminosity of 4.2 fb^{-1} , and 38.5 fb^{-1} was delivered by the LHC to the ATLAS detector in 2015 and 2016, respectively. Run 2 will continue through 2018, after which two year shutdown will be used to repair and upgrade the LHC and its experiments in anticipation of Run 3. The harsh radiation environment, high detector occupancies, and rate limitations are the reasons for long shutdowns. The ATLAS collaboration foresees three main upgrade projects for the upcoming shutdown: a replacement of the first endcap station of the New Small Wheel [82], higher-granularity for the existing calorimeter trigger [83], and the Fast Tracker (FTK) which is being commissioned into the ATLAS trigger system and will be briefly discussed in this chapter.

4.2 Luminosity and Pileup

The expected number of events N_i for a process (i) with a production cross section of σ_i is given in terms of the instantaneous luminosity (\mathcal{L}) as

$$N_i = \sigma_i \int \mathcal{L} dt. \quad (4.1)$$

The instantaneous luminosity can be expressed as a function of the rate of pp interactions and in terms of beam parameters as

$$\mathcal{L} = \frac{N_b^2 n_b f_{\text{rev}}}{4\pi\sigma_x\sigma_y} F, \quad (4.2)$$

where N_b is the number of protons per bunch, n_b is the number of bunches injected at the LHC per revolution, f_{rev} is the machine revolution frequency which is approximately 11 kHz, σ_x and σ_y stand for the horizontal (x -scan) and vertical (y -scan) Gaussian widths

Parameter	2010 – 2011	2012 – 2013	2015	2016
Beam energy (TeV)	3.5	4	6.5	6.5
Bunch spacing (ns)	50	50	50 – 25	25
Max number of bunches (n_b)	1380	1380	2244	2200
Protons per bunch (N_b)(10^{11})	1.45	1.6	1.15	1.15
Peak luminosity (10^{33} cm $^{-2}$ s $^{-1}$)	3.7	7.7	5.0	13.6
Integrated luminosity (cm $^{-2}$)	5.46	22.8	4.2	38.5
Mean pileup	9.1	20.7	13.7	24.2

Table 4.1: Operating parameters of the LHC for each data taking period [84, 85].

of the colliding beams, and F is the geometric luminosity reduction factor that serves as a small correction factor to account for the crossing angle between beams at the interaction point. The LHC design luminosity is $\mathcal{L} = 10^{34}$ cm $^{-2}$ s $^{-1}$. Table 4.1 presents the different parameters of the LHC for each data taking period until the end of 2016.

Due to the high frequency of collisions and the high density of the beam bunches, many pp interactions may occur simultaneously, called pileup. They result in the overlap of the electronic signals from multiple interactions and are categorized as *in-time pileup* or *out-of-time pileup*. *In-time pileup* events are caused by multiple pp interactions in the same bunch crossing, while *out-of-time pileup* occurs when traces from an event in a different bunch crossing are recorded. Increasing N_b or n_b , in Equation 4.2, results in higher luminosity but also raises the level of pileup. Higher N_b produces more interactions within a given bunch crossing, meaning higher *in-time pileup*. Large n_b reduces the bunch spacing, causing interactions from different bunch crossings to overlap (*out-of-time pileup*). The average number of interactions per bunch crossing in the 2016 dataset was found to be $\langle \mu \rangle \sim 24.9$ [86].

4.3 The ATLAS Detector

ATLAS is a general purpose detector located at Interaction Point 1 on the LHC ring. It is a hermetic detector of nearly 4π radians of solid angle coverage around the central interaction point, which is essential for reconstructing the energy flow in an event. The ATLAS detector is the largest volume particle detector ever constructed. It weighs approximately 7000 tons and has a cylindrical profile, 25 m in diameter and 44 m in length. It consists of a series of concentric cylinders around the interaction point where the proton and ion beams of the LHC collide. The ATLAS detector is composed of four major

components, as illustrated in Figure 4.2. Charged particle tracks are reconstructed in the Inner Detector (ID), composed of three sub-detectors as detailed in Section 4.3.1. The ID is enclosed by a thin solenoid, providing an axial magnetic field of 2 T that bends the trajectory of charged particles, allowing the measurement of their momenta. Charged and neutral particles exiting the ID are absorbed and measured in the sample electromagnetic (EM) and hadronic calorimeters, as explained in Section 4.3.2. The muon spectrometer, described in Section 4.3.3, surrounds the ATLAS calorimeters and measures the position and energy of charged muon tracks. The muon spectrometer is surrounded by three large air-core toroids, and the presence of the magnetic field allows the measurement of muon momenta.

ATLAS uses a right-handed coordinate system, as shown by the red and blue lines in Figure 4.2, with its origin at the nominal interaction point in the center of the detector; the z -axis is along the beam pipe, and the x - y plane is transverse to the beam direction. The positive x -axis is defined as pointing from the interaction point towards the centre of the LHC ring and the positive y -axis is defined as pointing upwards. Cylindrical coordinates (r, ϕ) are used in the transverse plane, in which ϕ is the azimuthal angle around the z -axis. The pseudorapidity is defined in terms of the longitudinal angle θ as $\eta = -\ln \tan(\theta/2)$, where η is a measure of the longitudinal angle against the beam line. A large value of η which is close to the beam line, is referred to as forward. The angular separation of two particles emerging from the interaction point is measured in units of $\Delta R \equiv \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$.

4.3.1 The Inner Detector

The ATLAS Inner Detector (ID) provides precision tracking of charged particles of ($p_T > 0.1$ GeV) with high efficiency over the pseudorapidity range of $|\eta| < 2.5$. The ID consists of three independent sub-systems at various radii between 3.3 and 101.6 cm away from the beam axis. Figure 4.3 shows the arrangement and radial distance of the barrel ID components. The ID is immersed in a uniform 2 T axial magnetic field generated by the central superconducting solenoid held at 4.5 K by liquid helium in the region of $|\eta| < 1.6$. The solenoid consists of superconducting NbTi cables and a light weight aluminum cylinder, reducing the amount of non-active material in the detector. The strong magnetic field

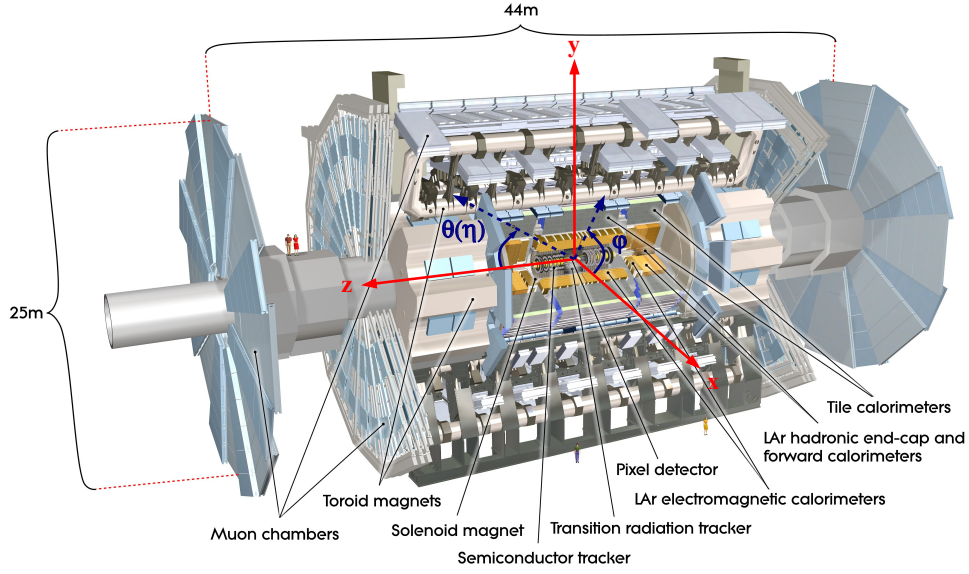


Figure 4.2: A computer generated image of the ATLAS detector showing the Inner Detector (Pixel, SCT, TRT), the Liquid Argon Calorimeter, the Tile Calorimeter, the Muon Detectors, and the toroid and solenoid magnets [87]. The ATLAS coordinate system is indicated by the red and blue lines (see text for more details).

deflects and bends charged particles within the ID, allowing their momenta to be accurately measured using the curvature of their tracks. Each sub-detector is split into cylindrical concentric barrel modules covering the central region and disk-shaped end-cap modules covering the forward regions. The tracks in the ID are reconstructed from individual hits, from many layers of the different systems: The Insertable B-Layer (IBL) and the Pixel detector, being closest to the Interaction Point, mostly contribute to the vertex finding of secondary vertices, the silicon microstrip (SCT) enhances moment resolution by adding higher radius hits, and the transition radiation tracker (TRT) enhances the particle identification through the pattern recognition and improves the momentum resolution recording an average of 36 hits per track.

Pixel Detector and IBL

The Pixel detector [89] surrounds the beam pipe and is the closest sub-system to the beam pipe. Due to this geometry it has the highest particle fluxes, which require

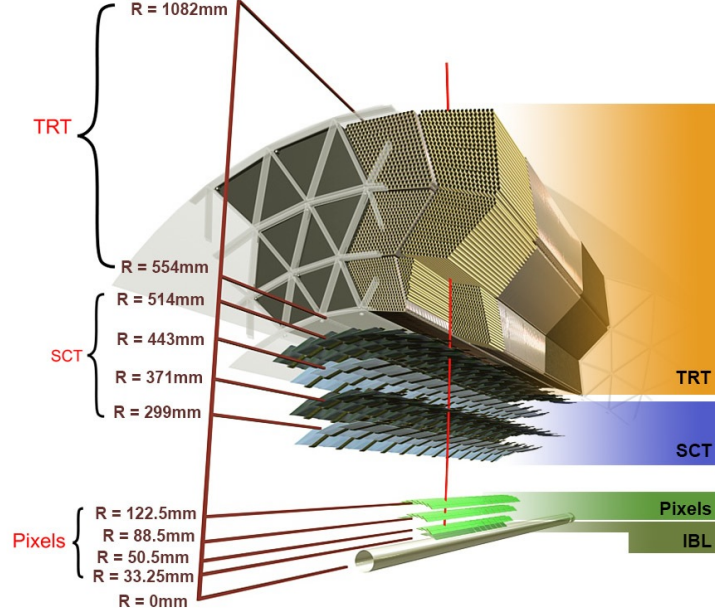


Figure 4.3: A cartoon of the central region barrel of the ATLAS Inner Detector showing the Pixel layer (including IBL at 33.25 mm), SCT, and TRT [88]. Distances of concentric layers from the beam axis are drawn to scale and labeled.

the highest granularity. The pixel detector consists of 4 barrel and 2×3 end-cap layers of silicon semiconductor pixel sensors that locate spatial hits and measure the energy deposited by ionizing particles. The barrel region has a coverage between 3.3 cm and 15 cm, and each end-cap consists of three disks with a coverage of $|\eta| < 2.5$. Each of the pixels is a reverse-biased p-n junction that is sensitive to incident charged particles, which create an electron-hole pair, both of which drift to the respective electrodes and induce electrical signals, which are then read out using about 80 million channels. Each pixel is of $50 \times 400 \mu\text{m}^2$ in area and $250 \mu\text{m}$ thick. This provides a spatial hit resolution for a point on a charged particle's trajectory of $10 \mu\text{m}$ in the r - ϕ plane and $115 \mu\text{m}$ along z and r .

For Run 2, a fourth innermost layer was installed in the barrel region in 2014, the insertable B-layer (IBL) [90]. The IBL is at a radial distance of 3.3 cm from the beam pipe and provides additional 8 million pixels over 12 ϕ sectors. Each pixel provides a spatial hit resolution of $8 \mu\text{m}$ in r - ϕ plane and $40 \mu\text{m}$ along the z axis. The addition of the IBL improved the track reconstruction, provided more precise vertex measurement and identification of jets originating from b -quarks, which typically decay beyond this radius. The improvement on the performance of the identification of b -jets, referred to as

b -tagging, due to the addition of the IBL is about 10% [91].

Semiconducting Tracker

The Semiconducting Tracker (SCT) lies outside the Pixel detector and is also made of silicon semiconductor sensors. Instead of pixels, the SCT sensors are segmented into strips with a pitch of $80\text{ }\mu\text{m}$. The SCT consists of four cylindrical layers in the barrel region and nine disks at each end of the barrel (endcap). Two individual layers of strips are closely laid at a slight angle of $\pm 20\text{ mrad}$ around the geometrical centre of the sensors to form a double layer. This creates a stereo-pairing which results in an improved spatial resolution along the strip length in the z direction. The four double-layers of silicon strip modules in the barrel regions are aligned parallel to the beam axis and covering radii between 29.9 and 56.0 cm. While the nine disks of double layer strips in the end-cap region extend a coverage up to $|\eta| < 2.5$. A hit along the strip has an intrinsic resolution of $17\text{ }\mu\text{m}$ in $r\phi$ and $580\text{ }\mu\text{m}$ in z and r .

Transition Radiation Tracker

The SCT is followed by the Transition Radiation Tracker (TRT) in radial direction. The TRT is composed of straw tubes measuring 144 cm in length and 4 mm in diameter. The straws are operated with a gas mixture of 70% Xe, 27% CO₂, and 3% O₂. The center of each straw tube has a single gold-coated wire, which is $31\text{ }\mu\text{m}$ in diameter and serves as an anode kept at ground potential. The walls of each straw serve as a cathode and are kept at a negative potential of approximately -1.5 kV . When charged particles traverse the straw, they ionize the gas mixture, causing electrons and positive ions to drift apart in the electric field and producing a detectable signal proportional to the energy deposited by the particle. A typical track passes through about 30 straws and the combined information yields a spatial hit resolution of $130\text{ }\mu\text{m}$ in a plane perpendicular to the wire. The TRT does not provide tracking information in the direction parallel to the straws.

The barrel region consists of 72 layers of 144 cm long straw tubes which cover a radius of 56.3 to 106.6 cm and are parallel to the beam axis. The end-cap regions consist of 160 layers of 36 cm long tubes which are radially oriented on 18 wheels. A total of 350848

straw tubes are used to improve the tracking resolution up to $|\eta| < 2.5$.

The layers of straws are separated by a polypropylene radiator which change the refractive index of the volume and provide discrimination between electrons and heavier charged particles. Electrons passing through the radiator will release a notably larger amount of transition radiation than heavier charged particles, such as pions. Therefore, the TRT plays an important role in electron identification and provides substantial discriminating power between electron and pions over the energy range between 1 and 200 GeV.

A typical charged particle of $p_T > 0.5$ GeV traversing the ID barrel will produce 4 pixel hits, 8 SCT hits and more than 30 TRT straw hits. The hits from all layers of the ID are combined into a single particle track using track finding algorithms.

4.3.2 Calorimeter

Particles exiting the ID are stopped in the ATLAS calorimeters [92] to measure their energies. The calorimeters cover the pseudorapidity range up to $|\eta| < 4.9$, and are segmented into towers in both η and ϕ , pointing towards the center of the detector. An overview of the ATLAS calorimeter system is illustrated in Figure 4.4. Sampling calorimeters [93] are used, consisting of alternating layers of dense passive material and an active medium. Particles passing through the passive material induce particle showers in which one particle produces a cascade of secondary particles of lower momentum, inducing a signal in the active medium through ionization or scintillation that is proportional to the total released energy.

The active material is composed of plastic scintillators or liquid argon that reacts in the presence of charged particles. Electromagnetic particles interact with the plastic scintillators by exciting valence electrons whose de-excitation produces photons. The number of produced photons is proportional to the deposited energy of the incident particle. Incident charged particles passing through the liquid argon medium ionize the liquid, electrons and positive ions drift towards the electrodes that measure the deposited charge. The passive material is composed of heavy absorber material that interacts with charged and neutral particles but does not measure the deposited energy.

Incident particles interact with the material through various mechanisms. The in-

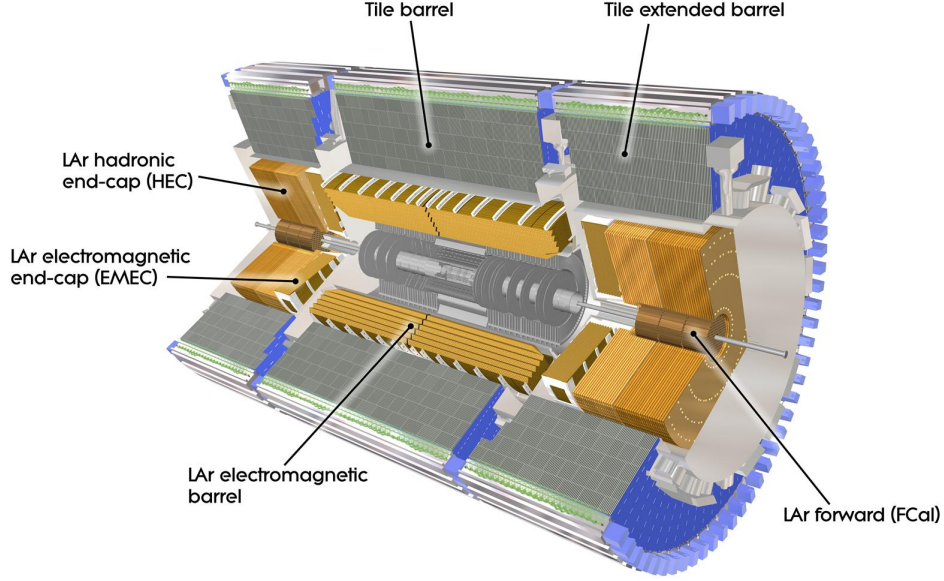


Figure 4.4: Image of the ATLAS detector to scale with a focus on the calorimeters, including the Tile barrel, Tile extended barrels, EMB, EMEC, HEC, and FCal [87].

teraction process of photons and electrons is characterized by the radiation length X_0 . Photons and high energetic electrons lose their energy via e^+e^- -pair production and bremsstrahlung respectively when passing through matter. Therefore, X_0 is equivalent to $7/9$ of the mean free path of a photon or the mean distance over which the electron loses all but $1/e$ of its energy¹. Hadrons typically lose their energy through inelastic hadronic collisions in matter, causing showers of particles. The mean free path of a hadron and the characteristic length of the hadronic showers is given by the nuclear interaction length λ . As a result, the calorimeters must be adequately large in order to fully capture interactions of various lengths of X_0 and λ , to precisely measure energies, and to avoid losing energy into the muon spectrometer. Muons deposit a small amount of energy in the calorimeters because they loose less energy through bremsstrahlung as they are more massive than electrons and they are not strongly interacting. Neutrinos do not interact with the calorimeters at all and appear as momentum imbalance inside the detector.

1. Here, e stands for the Euler's number and not the elementary electric charge.

The fractional calorimeter resolution as a function of energy is expressed as

$$\frac{\sigma_E}{E} = \frac{N}{E} \oplus \frac{S}{\sqrt{E}} \oplus C, \quad (4.3)$$

where N stands for the measurement of the noise due to background and electronics which is dominant at low energies, S parameterizes the stochastic uncertainty caused by the random sampling nature, and C is a constant term that reflects the non-uniformities in the detector and is dominant at higher energies. These terms are added in quadrature (\oplus) to obtain the fractional resolution.

ATLAS has two separate calorimeter systems. The Electromagnetic calorimeters, which measure the energy of electrons and photons, and the Hadronic calorimeters, which measure the energy of strongly interacting particles.

Electromagnetic Calorimeter

The EM calorimeter is used to detect electrons and photons within $|\eta| < 3.2$ with a gap at $1.37 < |\eta| < 1.52$ where the barrel components stop and the end-cap starts. They are the closest calorimeter to the interaction point and are composed of alternating layers of lead absorber and liquid argon (LAr) active material. LAr is chosen as an active material because it is radiation-hard and offers an intrinsically linear response which is stable over time. The argon is held in a liquid state at 89 K through cryostats. The absorber plates and electrodes are arranged in an accordion-like geometry that ensures coverage in ϕ , as shown in Figure 4.5. When electrons and photons pass through the lead, electrons emit bremsstrahlung and photons convert to e^+e^- -pairs. A cascade of photon and e^+e^- conversions produces an electromagnetic shower which ionizes the LAr. The liberated electrons drift towards the electrodes and induce electrical signal which is then processed by the readout electronics. The EM calorimeter is segmented into cells of $\Delta\eta \times \Delta\phi = 0.003 \times 0.025$ with three layers in depth. This fine segmentation is important to distinguish single photons from $\pi^0 \rightarrow \gamma\gamma$ decays. The total thickness of the EM calorimeter varies between 22 and 33 X_0 , ensuring that the energy of electrons and photons are almost completely contained within the EM calorimeter. The response resolution of the stochastic

and constant terms, given by Equation 4.3, were measured to be

$$\frac{\sigma_E}{E} = \frac{10\%}{\sqrt{E \text{ GeV}}} \oplus 1\%, \quad (4.4)$$

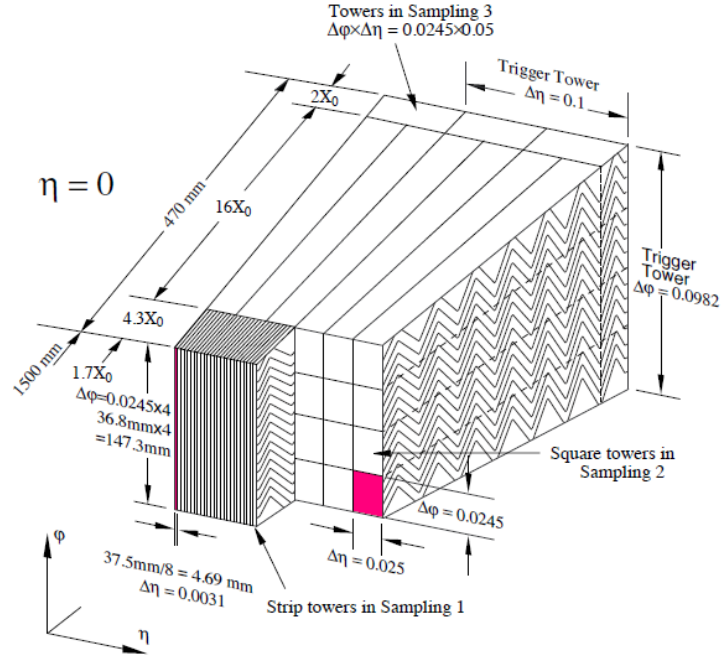


Figure 4.5: Sketch of a barrel module of the EM calorimeter, showing the different layers with their respective granularities in η and ϕ [87].

Hadronic Calorimeters

The hadronic calorimeters, which surround the EM calorimeters, contain and measure the energy of hadron showers within $|\eta| < 3.2$. The ATLAS hadronic calorimeters are divided into three parts: the Tile hadronic Calorimeter (TileCal), the liquid-argon Hadronic End-cap Calorimeter (HEC), and the liquid-argon Forward Calorimeter (FCal), as shown in Figure 4.5.

The TileCal consists of plastic polystyrene scintillator tiles with steel absorbers, and is made up of three parts: a barrel that covers the region up to $|\eta| < 1.0$, and two extended barrels on each side with a coverage range of $0.8 < |\eta| < 1.7$. The TileCal extends from an

inner radius of 2.28 m to an outer radius of 4.25 m. Each of its regions is segmented into 64 wedge-shaped modules in ϕ that contain the scintillator, steel and read-out electronics. The electronics are kept in steel support structures furthest from the beamline in order to reduce radiation exposure. The polystyrene and steel are oriented radially into three read-out layers that allow the measurement of longitudinal shower profiles. The polystyrene plates are connected to the photomultiplier tubes (PMTs) using wavelength shifting fibers. These convert the ultraviolet light produced in the scintillator due to passing charged particles into an amplified electrical signal. The TileCal contains about 4672 readout cells, each is read-out on both sides by two PMTs. This requires a total of 9852 PMTs to service all the detector.

In test beams [94] the response resolution to isolated charged pions of the combined LAr and tile calorimeter, expressed in Equation 4.3, of the stochastic and constant terms is

$$\frac{\sigma E}{E} = \frac{53\%}{\sqrt{E \text{ GeV}}} \oplus 3\%, \quad (4.5)$$

which is close to the design specifications.

The HEC, which has a coverage range of $1.5 < |\eta| < 3.2$, is based on the LAr technology. It uses copper instead of lead as a passive absorber material with a flat-plate design. The response resolution to isolated charged pions is

$$\frac{\sigma(E)}{E} = \frac{71\%}{\sqrt{E \text{ GeV}}} \oplus 1.5\%, \quad (4.6)$$

The FCal, which has a coverage range up to $|\eta| = 4.9$ uses copper as the absorber material for the first layer and tungsten for the second and third layers. The response resolution to isolated charged pions is

$$\frac{\sigma E}{E} = \frac{94\%}{\sqrt{E \text{ GeV}}} \oplus 7.5\%, \quad (4.7)$$

4.3.3 The Muon Spectrometer

The muon spectrometer (MS) surrounds the calorimeters and is the outermost sub-detector system of ATLAS. It is devoted to measure the momenta and position of muons that pass

through the ID and calorimeters. The MS system has approximately 1 million channels and extends from a radius of 5 m to 10 m with a small gap at $|\eta| = 0$ for service cables. The MS is composed of 3 concentric cylinders in the barrel region and is designed to measure the momentum of muons above 5 GeV and provides a resolution of 3% at 100 GeV. Four wheels cover the end-cap which extend the coverage up to $|\eta| < 2.7$. The spectrometer allows for precise tracking of muons which are bent by a large air-core toroid magnet system with a field between 0.5 and 1 T and allows an accurate measurement of muon momenta.

The muon chambers consist of two sets: one is dedicated to precision measurement of muon tracks and the second is dedicated to triggering on passing muons. Two types of precision chambers are used; the monitored drift tubes (MDT) [95] and cathode strip chambers (CSC) [96]. The MDT covers most of the MS pseudorapidity range of $|\eta| < 2.7$ except for the innermost layer of the endcap regions of $2.0 < |\eta| < 2.7$ where the CSCs are installed. The MDT consists of 3 cm diameter drift tubes which contain a mixture of 93% argon and 7% CO₂. Each tube has a single tungsten-rhenium wire that operates at a voltage of 3 kV and facilitates the measurement of the drift time of the ionization charge produced by incident particles. The typical spatial hit resolution of a single tube is below 100 μm and is improved to about 50 μm through the use of 3 or 4 layers of tubes in each chamber depending on its position in the detector.

The CSC are made up of multi-wire proportional chambers with orthogonal planar cathodes. They can resolve a higher occupancy and have a higher radiation tolerance than the MDTs and are therefore placed in the forward region of $2 < |\eta| < 2.7$, where the particle flux is larger. The radially oriented wires are held at a potential of 1.9 kV and are held at 2.5 mm away for each strip cathode. Typical tracking resolution obtained with the CSC detector in the bending plane is about 60 μm and has a high radiation tolerance and therefore is used as the first layer of the MS.

The precision chambers typically have a long charge collection time, about 700 ns for the MDT and 40 ns for the CSC. This large difference in the collection time is due to the differences in the design of the MDT and the CSC. The MDTs are tubes with a voltage applied on the central wire, where the field drops by $1/r^2$ (r is the radius) for points further away from the center. While, CSCs are flat chambers with a constant voltage difference and a constant field. Two dedicated trigger chambers provide fast measurements

for use in trigger decisions. The trigger system for muon events is based on Resistive Plate Chambers (RPC) [97] instrument in the barrel region of $|\eta| < 1.05$ while Thin Gap Chambers (TGC) [98] are used in the higher background environment of the endcap region up to $|\eta| < 2.4$. The RPC is composed of parallel electrode plates which are 2 mm apart and filled with a gas mixture of $\text{C}_2\text{H}_2\text{F}_4$. They are operated at a potential difference of 9.8 kV, allowing a very good timing resolution of about 2 ns. The TGC is composed of multi-wire proportional chambers with a gas mixture of CO_2 and $\text{n-C}_5\text{H}_{12}$. The anode wires of the TGC are held at 1.4 mm away from each strip cathode and held at a potential difference of 2.9 kV, allowing a timing resolution of about 4 ns.

The toroid magnets produce a magnetic field of 0.5 up to 1 T in the azimuthal plane that bends muons in the r - ϕ plane. There are eight rectangular coils in the barrel with a coverage of $|\eta| < 1.6$, and eight coils in each end-cap with a coverage of $1.4 < |\eta| < 2.7$. The coils are made up of a mixture of aluminum, copper, niobium, and titanium and are cooled with liquid helium to 4.5 K. The muon p_{T} resolution of the MS is limited by the non-uniformity of the magnetic field.

4.3.4 *The Trigger and Data Acquisition*

The high luminosity of the LHC produces numerous interactions per second while only a small fraction of them can be recorded due to the limitations in data storage capacity and rates. The ATLAS trigger system performs the run-time event selection, recognizes, and saves only the most interesting events after a sequential series of increasingly strict filters [99].

The current ATLAS trigger system consists of a hardware-based level (L1) trigger using coarse measurements from the calorimeters and muon systems, and a software-based High-level trigger (HLT). The L1 reduces the event rate from the bunch-crossing rate of 40 MHz to 100 kHz and the HLT further reduces it to an average recording rate of 1 kHz [99]. A schematic overview of the ATLAS Data Acquisition (TDAQ) system is shown in Figure 4.6.

The L1 trigger system performs the initial event selection and accepts events at a 100 kHz rate. It is optimized to provide a fast decision. It searches for high energy leptons, photons, and jets using a combination of information from the calorimeter and

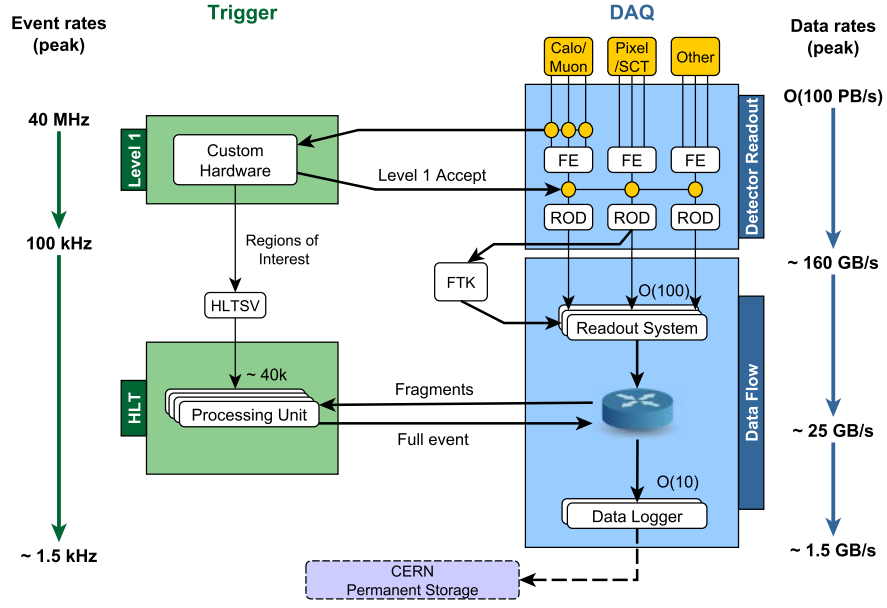


Figure 4.6: Schematic view of the ATLAS trigger system showing output rates in Run 2 [99].

MS. Electrons and photons are triggered on energy deposits in the EM calorimeter which is limited by its fine segmentation in $|\eta| < 2.5$. In the hadronic calorimeter, jet candidates are constructed at L1 from coarse calorimeter towers made of trigger-elements using a sliding window algorithm. A trigger-element is determined by the sum of cells in a 0.2×0.2 ($\eta - \phi$) region, and the sliding window examines the total E_T against a trigger threshold value in a 4×4 region of trigger-elements. Muon triggers are based on a coincidence of hits among several layers of the trigger chambers.

The L1 is followed by the HLT which operates at 1 kHz. The HLT consists of the Level 2 (L2) trigger followed by the event filter (EF). The L2 trigger performs similar measurements as the L1 trigger, but with a finer granularity and ID measurements for regions of interest. The EF fully reconstructs the event using offline tracking and jet reconstruction. The event reconstruction is performed using the ATLAS Athena control framework [100]. Most events that pass the selection requirements of the EF are written to the "main" analysis stream, while a few events which require a longer processing time are saved to a "debug" stream for reprocessing. Data is periodically reprocessed to reflect

software updates and increased understanding of the detector conditions. Recorded events are checked for data quality and those recorded during periods of sub-detector malfunction are flagged for removal from analysis.

Figure 4.6 shows the Fast Tracker (FTK) system that receives input from the ATLAS silicon tracking detectors after each L1 trigger and provides full-event track information to the HLT. FTK is being commissioned into the current ATLAS trigger system and a brief overview is given in the following section.

4.3.5 *Fast TracKer*

The increase in the average number of collisions per bunch crossing and the higher detector occupancy will create a busy and challenging environment for data readout and particle reconstruction. Limited computing resources will require the online data processing to reduce the data output to storage to a manageable level, and sophisticated trigger algorithms will be essential for selecting the events with interesting physics signatures, such as b -jets and τ -leptons, at a high efficiency while rejecting an increasingly large background. Therefore, the ATLAS trigger system needs to make better use of the information received by the silicon detector in order to improve charged particle reconstruction in the heavy pileup environment. As a consequence, the Fast TracKer (FTK) [101] is being developed to be included within the L1 and the HLT systems and designed to perform full track reconstruction of the complete granularity of the ID.

The FTK system is designed to run before the HLT for every event passing the L1 trigger at 100 kHz with an average latency of about 100 μ s. It will receive data from 98 million channels, and it will reconstruct trajectories in the silicon detector for charged particles with a transverse momentum above 1 GeV and within $|\eta| < 2.5$. At the end, it will provide tracks, reducing the need of tracking in the HLT.

The FTK algorithm is composed of two main steps. The first step is a pattern recognition in the Associative Memory (AM) for coarsely locating track candidates. Patterns are evolved using hits from 84 of the 12 detector layers; 2 of the 4 pixel layers and 6 of the 8 axial and stereo channels from the SCT. Potential patterns are pre-calculated using Monte Carlo simulation and stored for reference in a Pattern Bank. Hits in each event are compared with all the patterns in the Pattern Bank and track candidate are found, called

roads.

The second step depends on the Track Fitter for fitting the full-resolution hits in each candidate road to determine the optimal track parameters and reject false pattern matches. Track parameters are evaluated using the following linear combination:

$$p_i = \sum_j a_{i,j} x_j + b_i, \quad (4.8)$$

where p_i is a helix parameter or a term used in the χ^2 fit, $a_{i,j}$ and b_i are pre-calculated constants from MC simulations, and x_j stands for the hit coordinates in the silicon layers. Each road is attributed to a "sector" in which fixed $a_{i,j}$ and b_i are valid.

The FTK algorithms are implemented in electronics boards using VME and ATCA² standards. The final system will have about approximately 2000 FPGAs and 8000 Associative Memory (AM) custom ASICs. The design of the FTK system allows to rapidly carry out what is usually the most CPU intensive aspect of tracking, presently performed mostly by the HLT, by employing massive parallelism as the data pass through FTK. The system starts with 32 Data Formatter (DF) boards, which receive pixel and silicon strip data. The DF performs dedicated cluster-finding algorithms on pixel hits, and reorganizes the data, combining hits into η - ϕ towers, and sends them to the 128 associative memory boards (AMB) and auxiliary cards (AUX), to be processed. The AM board and AUX cards perform the pattern matching and the first stage fitting. Then, the data is sent to the 32 second stage boards (SSB) where the candidate roads that passed the first stage fits are supplemented with the cluster centroids from the 4 unused layers and a second stage fit is performed. Duplicate tracks are removed in the SSB before being sent to the 2 FTK to Level-2 interface cards (FLIC) boards. The FLIC organizes the final tracks from the SSB, reformat them, and send them to the HLT Readout System (ROS).

The system is being commissioned towards taking data in 2018. An example of the performance of the FLIC boards, which are installed in the ATLAS detector cavern, is illustrated in Figure 4.7. The event rate is shown as a function of the number of tracks per event record. The FLIC boards have been tested for all requirements and demonstrated to perform at or above the design threshold of 100 kHz needed by the HLT.

FTK will enhance the ATLAS trigger system and will provide the power of tracking

2. ATCA stands for Advanced Telecommunications Computing Architecture.

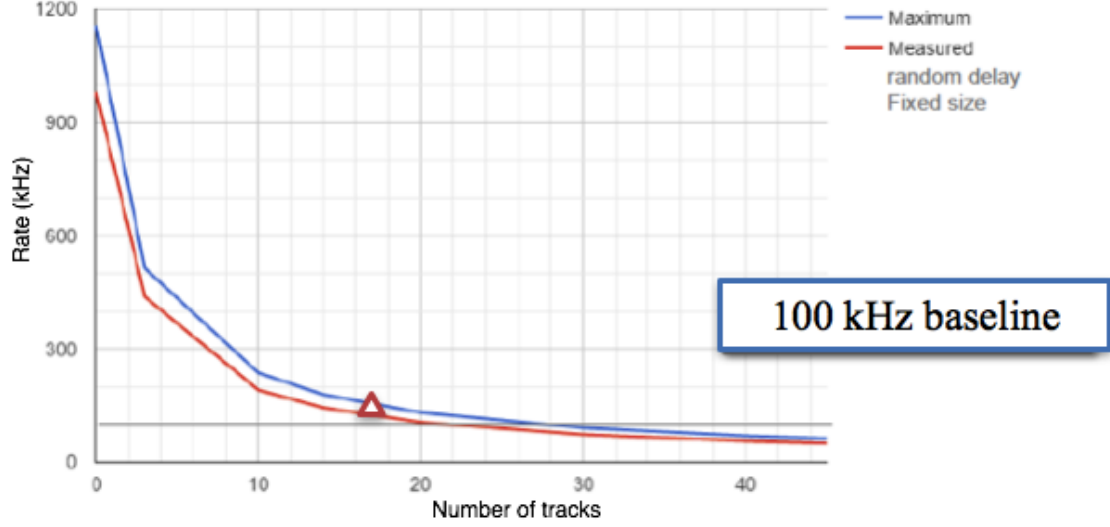


Figure 4.7: The measured (red line) and the expected (blue line) rate of events sending of the FLIC to HLT as a function of the number of tracks per event record. Events were sent with a random delay between them at a fixed rate. The red triangle shows the FTK specification which is 100 kHz of 17 tracks per event record.

after L1 trigger. When the LHC luminosity increases, the track availability from FTK will improve the performance of particle identification. Further information on FTK can be found in the proceedings that the author published in [102].

4.3.6 Luminosity Measurement

The measurement of the beam luminosity is needed to determine cross sections of observed process. The two main dedicated detectors to monitor the bunch-by-bunch luminosity are BCM (Beam Conditions Monitor) and LUCID (Luminosity measurement using Cherenkov Integrating Detector) [103].

BCM consists of four small diamond sensors arranged in a cross pattern, at a distance of 1.84 m corresponding to $|\eta| = 4.2$ [104], on each side surrounding Interaction point 1 and close to the beam pipe. In addition to luminosity measurements, BCM monitors the stability of the LHC beam.

LUCID is a Cherenkov detector located on each side of Interaction point 1, at a distance of 17 m corresponding to $|\eta| = 5.8$. It has sixteen mechanically polished aluminum tubes

filled with C_4F_{10} gas around the vacuum chamber.

Alternative measurements of luminosity can be provided by ALFA (Absolute Luminosity For ATLAS) [105], which is designed to determine the total pp cross section by measuring elastic scattering at very small angles in special runs with low beam divergence. The ALFA detector is placed at 240 m from the interaction point and is composed of scintillating fibre trackers. It detects elastic scattering at very small angles of $3 \mu\text{rad}$. At these small angles, the scattering amplitude relates to the total cross section by the optical theorem [106]. The ALFA detector aims at using Coulomb scattering [107], by fitting Equation 4.9, and thus determining the luminosity (\mathcal{L}), the parameter ρ , the total cross section (σ_{tot}) and the slope parameter b .

$$\frac{dN}{dt} = \mathcal{L} \cdot \pi \cdot |A_C + A_N|^2 \approx \mathcal{L} \cdot \pi \cdot \left| -\frac{2 \cdot \alpha_{EM}}{|t|} + \frac{\sigma_{\text{tot}}}{4 \cdot \pi} (i + \rho) \cdot e^{\frac{-b \cdot |t|}{2}} \right|^2, \quad (4.9)$$

where A_C is Coulomb interaction amplitude, A_N is the strong interaction amplitude, and α_{EM} is the electromagnetic coupling constant [107]. Additional cross checks on the luminosity measurement are provided by the Meddipix2 sensors [108] and the calorimeters, measuring the overall radiation at various points within the ATLAS detector.

The method for calibrating the ATLAS luminosity scale is based on the beam displacement technique known as the van der Meer (vdM) scans method (sometimes referred to as beam-separation or luminosity scans) [109]. The main idea of the vdM scans is to measure the effective convolved beams widths in dedicated fills during which beams are stepwise separated.

Chapter 5

DEFINITION OF PHYSICS OBJECTS

The search of $t\bar{t}H(H \rightarrow b\bar{b})$ involves the reconstruction and identification of electrons, muons, missing transverse energy, and jets, as well as the identification of b -jets. This chapter describes the reconstruction of particle tracks and particle energies from the signals in the ATLAS detector described before. Section 5.1 illustrates the characteristic of a charged particle track, and the reconstruction of primary vertices. Section 5.2 details the reconstruction, identification and isolation of electrons and muons. Section 5.3 summarizes the reconstruction algorithms and the calibration methods of jets. Section 5.4 details the b -tagging algorithm used to identify b -jets. Section 5.5 describes the missing transverse energy. Moreover, a brief description of the associated systematic uncertainties is presented.

5.1 Tracks and Primary Vertices

The tracking [110, 111] and vertexing [112] algorithms are both based on the inner detector information. A charged particle in the sub-detectors generates hits in the different layers which are later combined to obtain a track. In the ATLAS coordinate system, the helices produced by tracks in the magnetic field are characterized by five parameters to exploit the full geometry and kinematics of the incoming particles. A reconstructed track is fully characterized using the following set of parameters:

$$(d_0, z_0, \phi, \theta, q/|\vec{p}|), \quad (5.1)$$

where d_0 , and z_0 represent the track impact parameters in the transverse and longitudinal planes respectively, ϕ and θ express the azimuthal and polar angle respectively, and $q/|\vec{p}|$ stands for the charge over momentum. The impact parameter and the direction are usually expressed with respect to the reconstructed hard-scatter primary vertex in the event. Figure 5.1 illustrates a geometric definition of the track parameters.

The hits belonging to a track are found using an *inside-out* pattern recognition algorithm [110]. Meaning, the track finding starts building track *seeds* from space points in the silicon detectors, performs a first track reconstruction which is extended outwards to the TRT. Also an *outside-in* sequence known as back-tracking is used to take into

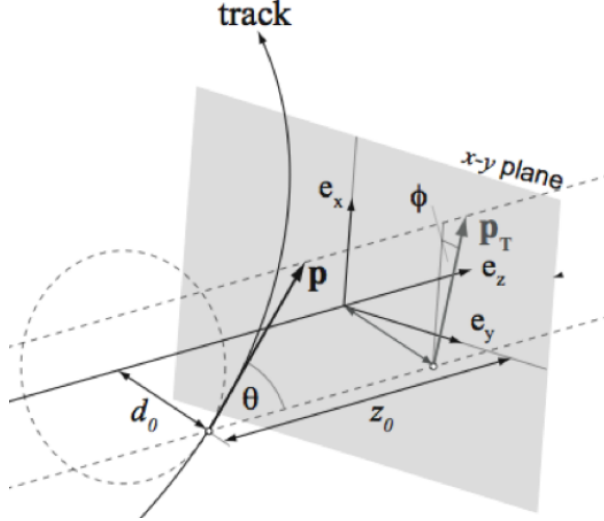


Figure 5.1: A geometric illustration of the ATLAS track parameterization (the parameters are defined in Section 5.1).

account all the hits that were not chosen in the previous algorithm. It is seeded in the TRT and then from the selected hits a track is formed, parametrized, and extrapolated to the silicon detectors.

Primary Vertices (PV) are reconstructed from the combination of reconstructed tracks with an adaptive vertex fitting algorithm [113] and are required to lie within the estimated position of the beam spot¹. Two steps are used in the reconstruction of the PVs: the primary-vertex finding where reconstructed tracks are associated to the vertex candidates, and the vertex fitting where the vertex position and the corresponding uncertainties are estimated. In order to enhance the resolution on the vertex spatial position, only vertices that have at least two tracks with a $p_T > 400$ MeV associated with them are considered.

The presence of pileup increases the number of reconstructed interaction vertices in the event. The vertex that has the highest sum of the squared track p_T is considered to correspond to the hardest pp interaction and is defined as the main vertex of the event. The rest of the PVs are then assumed to be pile-up interactions. Vertices which are incompatible with the beam collision region are considered as secondary vertices and will be discussed in Section 5.4.

1. The beam spot is referred to as the spatial region around the interaction point where the profiles of the two beams overlap.

5.2 Leptons

In the following, the reconstruction and identification of electrons and muons is discussed. Isolation is a crucial element to distinguish leptons from jets and it will be described as well. τ -leptons are not explicitly used in this thesis and therefore their reconstruction techniques are not discussed.

5.2.1 Electrons

Electrons are reconstructed in the central region of the ATLAS detector within $|\eta| < 2.47$, but outside the transition region ($1.37 < |\eta| < 1.52$) between the barrel and the end-cap EM calorimeter. Figure 5.2 illustrates an electron traversing elements of the ATLAS detector. An electron typically has 12 silicon measurement points (hits); starting with the IBL pixel layer, 3 pixel layers, and 4 double-sided silicon strips, and approximately 72 TRT layers. Then, the electron deposits its energy in four successive electromagnetic calorimeter layers listed according to the electron's trajectory: the presampler, a layer finely segmented in η (strips), a layer of roughly 16 radiation lengths and a backplane layer. This leaves about 2% of the electron's energy to reach the hadronic calorimeter.

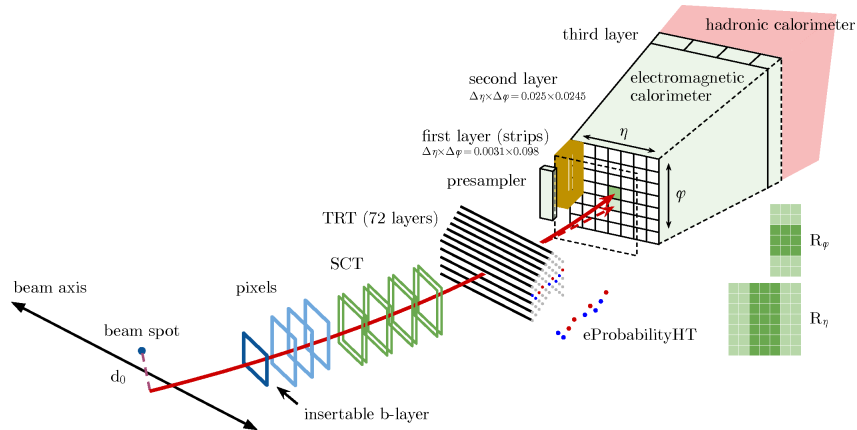


Figure 5.2: A schematic view of an electron (indicated by the black line) traversing elements of the ATLAS detector detailing the stages of the electron reconstruction and identification.

The first step in reconstructing an electron is the construction of clusters in the calorimeter energy deposits. Electrons make use of fixed-size clusters built with 3×5 cells

in angular units of 0.025×0.025 with respect to (η, ϕ) -space, where the size is fixed by the largest (middle) electromagnetic calorimeter layer [114]. Regardless of the size fixed by the central layer, all three layers are summed together to evaluate the total transverse energy deposit within this window. The window position is adjusted until the transverse energy is a maximum, which is referred to as the sliding-window algorithm [115]. If the combined transverse energy of the cluster is above 2.5 GeV, the region is marked as a seed. This threshold was chosen to optimize the reconstruction efficiency while minimizing the contribution from electronic or pileup noise [114].

Once the seed clusters are found, an attempt to match them to well-reconstructed tracks in the ID is made. If the matching fails, the cluster is tagged as an unconverted photon. If matching is possible and the track is not a primary vertex, then the cluster is tagged as a converted photon². Note that about 30% to 35% of identified photons are converted photons [116]. Finally, if matching is possible and the track comes from the hard-scatter vertex, then the cluster is tagged as an electron [114]. In favor of reducing the background arising from conversions and secondary particles, additional requirements on the track parameters are imposed, $|d_0|/\sigma_{d_0} < 5$ and $|z_0 \sin\theta| < 0.5$ mm, defined in Section 5.1. In the last stage of the electron reconstruction, the electron clusters are enlarged to 3×7 units in the barrel and 5×5 units in the endcap region of the calorimeter [115]. The clusters are specifically enlarged in the ϕ direction in order to capture the full electron energy including the lost energy from bremsstrahlung.

The reconstruction efficiency of electrons is defined as the ratio of the number of clusters matched to a track after passing the track quality criteria to the number of all clusters. The efficiency measurement of the electron reconstruction is based on the tag-and-probe method using the Z and the J/ψ resonances, as described in [115]. The tag-and-probe method is a data-driven technique which exploits well known resonances such as the Z boson, as a source for the production of electron-positron pairs. It selects events with a Z candidate using tighter requirements on the "tag" electron and looser requirements on the "probe" electron. The fraction of probe electrons which pass the selection under study gives an estimate of the corresponding efficiency. The reconstruction efficiency is found to have a mild dependence on the transverse energy of the electron (E_T), with values ranging

2. Photon conversions are processes in which a photon splits into an e^+e^- pair when interacting with the detector material.

from 97% for $E_T = 15$ GeV up to 99% for $E_T > 50$ GeV.

A set of requirements is used to distinguish signal electrons originating from the hard interaction *prompt* from other *non-signal* charged particles which have similar properties. For example, the misidentification of a charged pion as an electron can occur since a pion may leave an electron-like track in the inner detector. Another example, comes from photon conversions into pairs $e + e^-$ that happen in the detector leaving both tracks and energy deposit in the electromagnetic calorimeter that are often very difficult to distinguish from signal electrons and could be mistaken for an electron.

As a result, the electron identification is based on a set of calorimeter-based and track-based variables, defined in Table 5.1. This can be performed by imposing independent requirements on the discriminating variables, referred to as cut-based identification, or a single requirement on the ratio of the signal and background likelihood functions. The input to the likelihood functions are defined in Table 5.1, and known as likelihood-based (LH) identification³. The likelihood-based identification used in this thesis is based on a multivariate analysis technique and provides higher rejection of non-signal electrons for the same identification efficiency compared to the cut-based identification.

3. The LH method uses the signal and background probability density functions (PDFs) of the discriminating variables. An overall probability which is based on these PDFs and defined as the product of the individual PDFs, is calculated for the object to be signal or background.

Type	Name	Description
Hadronic leakage	R_{had1}	Ratio of E_T in the first layer of the hadronic calorimeter to E_T of the EM cluster (used over the range $ \eta < 0.8$ or $ \eta > 1.7$)
	R_{had}	Ratio of E_T in the hadronic calorimeter to E_T of the EM cluster (used over the range $0.8 < \eta < 1.7$)
Back layer of EM calorimeter	f_3	Ratio of the energy in the back layer to the total energy in the EM accordion calorimeter (used < 100 GeV)
Middle layer of EM calorimeter	$w_{\eta 2}$	Lateral shower width, $\sqrt{(\sum E_i \eta_i^2)/(\sum E_i) - ((\sum E_i \eta_i)/(\sum E_i))^2}$, where E_i is the energy and η_i is the pseudorapidity of cell i and the sum is calculated within a windows of 3×5 cells
	R_Φ	Ratio of the energy in 3×3 cells over the energy in 3×7 cells centered at the electron cluster position
	R_η	Ratio of the energy in 3×7 cells over the energy in 7×7 cells centered at the electron cluster position
Strip layer of EM calorimeter	w_{stot}	Shower width, $\sqrt{(\sum E_i (i - i_{\text{max}})^2)/(\sum E_i)}$, where i runs over all strips in a window of $\Delta\eta \times \Delta\phi \approx 0.0625 \times 0.02$, corresponding typically to 20 strips in η , and i_{max} is the index of the highest-energy strip
	E_{ratio}	Ratio of the energy difference between the largest and second largest energy deposits in the cluster over the sum of these energies
	f_1	Ratio of the energy in the strip layer to the total energy in the EM accordion calorimeter
Track conditions	n_{Blayer}	Number of hits in the innermost pixel layer (IBL)
	n_{Pixel}	Number of hits in the pixel detector
	n_{Si}	Number of total hits in the pixel and SCT detectors
	d_0	Transverse impact parameter with respect to the beam-line
	d_0/σ_{d_0}	Significance of transverse impact parameter defined as the ratio of d_0 and its uncertainty
	$\Delta p/p$	Momentum lost by the track between the perigee and the last measurement point divided by the original momentum
TRT	eProbabilityHT	Likelihood probability based on transition radiation in the TRT
Track-cluster matching	$\Delta\eta_1$	$\Delta\eta$ between the cluster position in the strip layer and the extrapolated track
	$\Delta\phi_2$	$\Delta\phi$ between the cluster position in the middle layer and the track extrapolated from the perigee
	$\Delta\phi_{\text{res}}$	Defined as $\Delta\phi_2$, but the track momentum is rescaled to the cluster energy before extrapolating the track from the perigee to the middle layer of the calorimeter
	E/p	Ratio of the cluster energy to the track momentum

Table 5.1: Discriminating variables used in the electron likelihood-based (LH) identification [117].

Three identification operating points were provided to identify prompt electrons and are listed here in order of increasing background rejection: Loose, Medium, and Tight. They are defined in such a way that each operating point uses the same variables to define the LH discriminant but with a different cut value. Therefore, electrons selected by TightLH are all selected by MediumLH and those selected by MediumLH are also selected by LooseLH. Electrons in this thesis are required to pass TightLH identification operating point. However, a looser identification operating point is used in the estimation of fake and non-prompt electrons as detailed in Chapter 6.

The identification efficiency is defined as the ratio of the number of electrons that pass the identification requirements to the total number of electron candidates. It is measured using the tag-and-probe method on $Z \rightarrow e^+e^-$ and $J/\psi \rightarrow e^+e^-$ events. The performance of the LH identification algorithm is illustrated in Figure 5.3. Depending on the operating point, the signal (background) efficiencies for electrons with $E_T = 25$ GeV are in the range of from 78 to 90% (0.3 to 0.8%) and increase (decrease) with E_T .

To further suppress the contribution from non-signal electrons, additional requirements on the total transverse momentum contained within a cone around the direction of the electron are imposed, the so called *isolation* requirements. Isolation requirements are based on track and calorimeter quantities. Two discriminating variables have been used: a track isolation, $p_T^{\text{varcone0.2}}$, which is defined as the sum of the transverse momenta of all the tracks within a cone of $\Delta R = \min(0.2, 10 \text{ GeV}/E_T)$ around the candidate electron track and originating from the PV of the hard collision. The second variable is the calorimetric isolation energy, $E_T^{\text{cone0.2}}$, which is defined as the sum of the transverse energies of the calorimetric cells in a cone of size $\Delta R = 0.2$ around the candidate electron. Electrons considered in this thesis must satisfy the *Gradient* isolation operating point. The isolation efficiency is defined as the ratio of the number of electrons passing a certain isolation selection to the total number of electron candidates passing the identification requirements. The *Gradient* operating point is defined so that the isolation efficiency is at least 90% for $p_T > 25$ GeV, increasing to 99% at 60 GeV [117].

Similarly to the reconstruction and identification efficiency, the electron isolation is measured using tag-and-probe method on $Z \rightarrow e^+e^-$ and $J/\Psi \rightarrow e^+e^-$ events. The low- E_T range (7 to 20 GeV) is covered by $J/\Psi \rightarrow e^+e^-$ events while $Z \rightarrow e^+e^-$ events are used for measurements above 15 GeV.

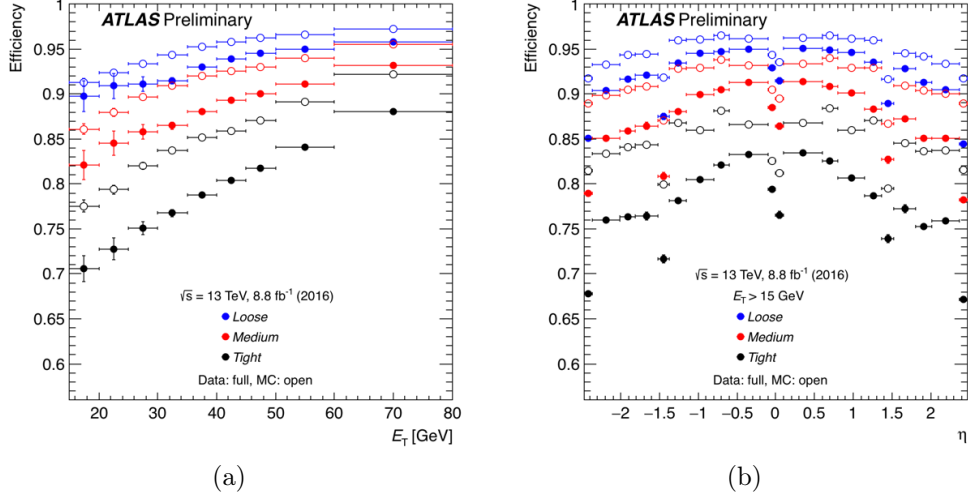


Figure 5.3: Electron identification efficiencies using tag-and-probe method on $Z \rightarrow e^+e^-$ events for the various operating points as function of (a) transverse energy E_T , integrated over the full pseudo-rapidity range and (b) pseudo-rapidity η for electrons with $E_T > 15$ GeV. The efficiencies have been measured using 8.8 fb $^{-1}$ of the 2016 data. The lower efficiency in data (full circles) than in simulation (open circles) arises from the fact that the simulation does not properly represent the 2016 TRT conditions, in addition to some mismodeling of the calorimeter shower shapes. Both of these differences between data and simulation were considered when optimising the likelihood-based selection criteria for 2016 data. The asymmetry near $\eta = 0$ seen in (b) is caused by the gap in the electromagnetic calorimeter, which is shifted by 2 mm in the z -direction with respect to the ATLAS reference frame [117].

Given the complexity of the electron reconstruction and identification requirements, it is expected that the detector simulation can only approximately describe the efficiency. Therefore, the simulated samples are corrected to reproduce the measured data efficiencies. The efficiencies are estimated in both data and in simulation and their ratio is used as a scale factor to correct the simulation. The scale factors which are measured as function of both η and E_T of the electron deviate from unity by few percent. The combined uncertainties on the scale factors are of the order of few percent at low E_T and below 1% at high E_T (above 30 GeV) [117].

5.2.2 Muons

Muons are reconstructed from tracks formed in the MS alone, or combining information from the MS with the ID. Different identification criteria define various muon "types". Four types of muons exist; Combined (CB) muons, Segment-tagged (ST) muons, Calorimeter-tagged (CT) muons and Extrapolated (ME) muons [118]. Muons used in this thesis are the CB muons and are discussed in the following.

Each of the three MDT muon spectrometer layers provides six to eight η measurements for a single muon passing through the detector within $|\eta| < 2.7$. Hits in each layer are combined to form local track segments, then the local track segments from each layer are combined to form an overall muon spectrometer track [118]. On the other hand, the inner detector provides an independent measurement of the muon trajectory close to the interaction point. A typical muon track within the acceptance of the inner detector has 3 pixel hits, 8 SCT hits, and 30 TRT hits (within $|\eta| < 1.9$) [118]. At the end, the algorithm uses tracks that are reconstructed independently in the ID and in the MS and performs a global refit, resulting in a combined track.

Muon identification is performed in order to disentangle prompt muons from background events mainly coming from pion and kaon decays. The muon identification uses the following set of variables:

- $|q/p|$ significance, defined as the absolute value of the difference between the ratio of the charge q determined from the track curvature and the momentum p of the muons measured in the ID and MS divided by the sum in quadrature of the corresponding uncertainties,
- ρ' , defined as the absolute value of the difference between the ratio of the transverse momentum measurements in the ID and MS divided by the p_T of the combined track
- the normalized χ^2 of the combined track fit [118].

In order to ensure a robust momentum measurement, additional requirements on the number of hits in the ID and MS are used. At least one pixel hit, at least five SCT hits,

fewer than three pixel or SCT holes⁴, and at least 10% of the TRT hits assigned to the track are included in the fit within $0.1 < |\eta| < 1.9$.

Muons in this thesis are required to pass the *Medium* identification criteria. This identification selection is the ATLAS standard selection which minimizes systematic uncertainties associated with the calibration and reconstruction of muons. Two additional requirements are needed for the *Medium* criteria: at least three hits in at least two MDT layers except in the $|\eta| < 0.1$ region, where tracks are allowed with at least one MDT layer but no more than one MDT hole layer, and a significance of $|q/p| < 7$ [118].

Similar to the electrons, in order to further reduce the contamination from non-prompt muons coming from the heavy-flavor hadron semi-leptonic decays, additional isolation requirements using a combination of variables from track-based and calorimeter-based are imposed. Muons considered in this thesis must satisfy the *Gradient* isolation operating point. The isolation efficiency is defined as the ratio of the number of muons passing a certain isolation selection to the total number of muons passing the *Medium* identification criteria. The *Gradient* operating point is defined so that the isolation is at least 90% for $p_T > 25$ GeV, increasing to 99% at 60 GeV [118].

Reconstruction, isolation, and identification efficiencies are measured in data and simulation using tag-and-probe method on $Z \rightarrow \mu^+\mu^-$ events for $p_T^\mu > 15$ GeV, and $J/\Psi \rightarrow \mu^+\mu^-$ for $5 < p_T^\mu < 15$ GeV events.

The muon momentum scale and resolution are studied in $Z \rightarrow \mu^+\mu^-$ and $J/\Psi \rightarrow \mu^+\mu^-$ events. Correction factors, as a function of the muon momentum in various η regions, are derived and applied to the simulated muon momentum to match the known value of the Z -boson mass.

5.3 Jets

Gluons and quarks created as final state partons of hadron collisions can not exist in isolation and are not directly observed in the detector, due to color confinement. Instead the strong color field between the partons causes a shower of additional gluons and quarks to radiate, which finally build color neutral objects, hadrons. Spray of collimated showers of hadrons are observed in the detector as jets. Jets are reconstructed from these

4. A hole is defined as an active sensor traversed by the track but it does not contain any hits.

particles and their energy deposits in the finely segmented calorimeter cells. The aim of jet reconstruction is to produce physics objects whose kinematics and characteristics are as close as possible to those of the initial partons.

5.3.1 Jet Reconstruction

Jets are reconstructed using clustering algorithms [119] that attempt to regroup the many particles of the spray of hadrons into a four-vector representing the energy of the initial hard scatter parton, and its direction. Jets may be defined in various ways depending on the type of objects and algorithms used to construct and build them. Particle jets are reconstructed from truth stable particles in MC samples, track jets are built from reconstructed tracks in the detector, and calorimeter jets (or simply jets) are built from energy deposits in the calorimeter. Jets in ATLAS are usually constructed from many topologically adjacent clusters of calorimeter cells called *topoclusters* as explained below.

A particle traversing the detector leaves energy deposits in the calorimeter cells that are grouped into a single topocluster. Topoclusters are formed through an iterative procedure [120] that starts with identifying the most significant energy deposits E_{cell} as the seed cells, and ends with clustering the neighboring cells into a single topocluster.

Jet finding algorithms [119] attempt to combine topoclusters, and decide which inputs are aggregated into individual jets. The energy measurement in a topocluster is assumed to be caused by a massless particle with four-vector of magnitude $E = \sum E_{\text{cell}}$ and directed from the center of the detector towards the energy-weighted barycenter of the topocluster. The jet finding algorithm groups the topoclusters that are likely to have resulted from the same initial parton together starting from the highest p_T topocluster referred to as the seed. Then, topological clusters within a radius R of the seed and satisfying specific requirements are grouped together. The distance between the four-vectors is defined as

$$\Delta_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2, \quad (5.2)$$

where y is the rapidity and ϕ is the azimuthal angle.

A new four-vector seed is created and nearby topoclusters within a radius R are recalculated at the energy-weighted barycenter of the grouped topoclusters as explained above. This is repeated until the energy-weighted barycenter of the grouped topoclusters

is fixed. Then, all grouped clusters in the event are replaced with a jet four-vector, and the whole procedure begins again by choosing a new seed four-vector with the highest p_T topocluster remaining in the event. At the end of this iterative procedure, all topoclusters should have been replaced by jets.

Jet finding algorithms have to be theoretically well defined at all orders in perturbation theory and the jet multiplicity should be insensitive to any modeling details of hadronization and parton showering. Therefore, they should be well-defined, collinear-safe, and infrared-safe [121], as explained below.

The boundaries of a jet should be well-defined even in the case where two jets would overlap. The algorithms will generally assign the shared topoclusters to one of the overlapping jets, depending on the energy and the distance between the topocluster and the other jet four-vectors.

Collinear-safe algorithms ensure that the formation of a jet is insensitive to the number of particles within the hadron shower, i.e. the jet boundaries should not be affected if a single particle is replaced by two collinear particles of half the original energy.

Lastly, the definition of the jet should be independent of the soft radiation of the initial parton. Infrared-safe algorithms require the jet clustering to be driven by the hardest energy deposits and ignore the low energy between overlapping jets.

The anti- k_T algorithm [122] successfully exhibit all the three requirements mentioned above. The algorithm combines the two four-vectors into a jet depending on the minimum p_T -weighted geometrical distance between them, as defined in equation 5.3. Moreover, the algorithm depends on the distance between each four vector and the LHC beam as defined by equation 5.4. In the below equations, k_{ti} is the p_T of input i , Δ_{ij} is the distance between inputs i and j as defined earlier in equation 5.2, R is a radius parameter that defines the size of the jet, and p is a configurable exponent.

$$d_{ij} = \min(k_{ti}^{2p}, k_{tj}^{2p}) \frac{\Delta_{ij}^2}{R^2}, \quad (5.3)$$

$$d_{iB} = k_{ti}^{2p}. \quad (5.4)$$

The jet finding algorithm proceeds by identifying the smallest of the distances d_{ij} between the two four-vectors of an event. If $d_{ij} < d_{iB}$ the two four-vectors are removed

and replaced by a single four-vector combination. Then, the smallest of the distances d_{ij} is recalculated and the sequential recombination procedure continues. If $d_{iB} < d_{ij}$ for all four-vector combinations, then the four-vector i is classified as a final jet and removed from the list of candidates. This sequential procedure continues until all inputs have been classified into jets and no four-vectors remain.

The choice of p in equations 5.3 and 5.4 defines the combination behavior of soft particles. Jets used in this thesis have a value of $p = -1$, which are referred to as anti- k_t jets and use a radius parameter $R = 0.4$, as shown in Figure 5.4.

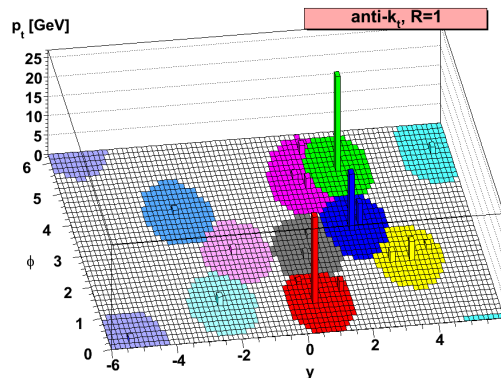


Figure 5.4: Illustration of topocluster grouping for the anti- k_t algorithm. The hard jets are all circular with a radius R , and only the softer jets have more complex shapes, the pair of jets near $\phi = 5$ and $y = 2$ provides an example in this respect [122].

5.3.2 Jet Calibration

The purpose of the jet calibration procedure is to correct the energy of the reconstructed jets in the detector to correspond to the energy of the initial parton at particle level. A series of corrections is derived from both MC simulation and data, the later referred to as the in-situ corrections [123]. The sequential calibration scheme for calorimeter jets is explained in the following.

Origin correction

The origin correction recalculates the jet four-vectors to point to the hard-scatter PV

rather than to the geometrical center of the detector as it was initially constructed. This correction improves the angular resolution of jets while only having a small effect on the jet p_T .

Pileup correction

The pileup correction takes into account additional energy deposited within the jet radius from in-time and out-of-time pileup. On average, the additional energy from pileup is deposited uniformly in η and ϕ through the detector causing a diffuse background that may be deducted from individual jets [123], [124]. The corrected p_T of an individual jet, p_T^{corr} is according to the following equation:

$$p_T^{\text{corr}} = p_T - \rho \cdot A - \alpha \cdot (N_{PV} - 1) - \beta \cdot \langle \mu \rangle, \quad (5.5)$$

where the level of pileup is parameterized as a function of the median energy density ρ of jets in the event, the number of primary vertices N_{PV} , and the average number of interactions per crossing $\langle \mu \rangle$. The pile-up energy is subtracted from each jet according to its area A . The jet area is defined using ghost association [125], where "ghost" particles of infinitesimal momentum are added uniformly to the event before jet reconstruction in order to probe the area assigned to the jet. The remaining terms in the above equation illustrate the residual corrections that remove the remaining effects for both in-time pileup $\alpha = \frac{\partial p_T}{\partial N_{PV}}$, and out-of-time $\beta = \frac{\partial p_T}{\partial \mu}$.

Jet Energy Scale and η Correction

The jet energy scale and η correction is derived from MC in order to correct the reconstructed jet energy at the electromagnetic (EM) scale to the true energy scale at particle level. It corrects the mismodeling due to unmeasured energy deposited in inactive detector regions and outside of the jet radius (out-of-cone radiation), reconstruction inefficiencies, and non-compensation of the hadronic calorimeters.

In order to derive this calibration, the true jet p_T is calculated in MC using isolated re-

constructed calorimeter jets that are matched geometrically to truth jets within $\Delta_{ij} = 0.3$. The ratio of reconstructed jet energy to the true jet energy is parameterized as a function of the reconstructed jet's p_T and η_{det} , and its inverse is applied as an energy correction. Note that the η_{det} is the η as measured towards the center of the detector, as opposed to the primary vertex, and which is useful when deriving average corrections that are geometrically dependent. Figure 5.5 shows the average energy response, which is the inverse of the jet calibration factor. Note the gaps and transitions between sub detectors of the calorimeter that result in a lower energy response which is evident when parameterized in η . These gaps are the result of absorbed or undetected particles [123]. Lower energetic jets need higher corrections, as shown in Figure 5.5.

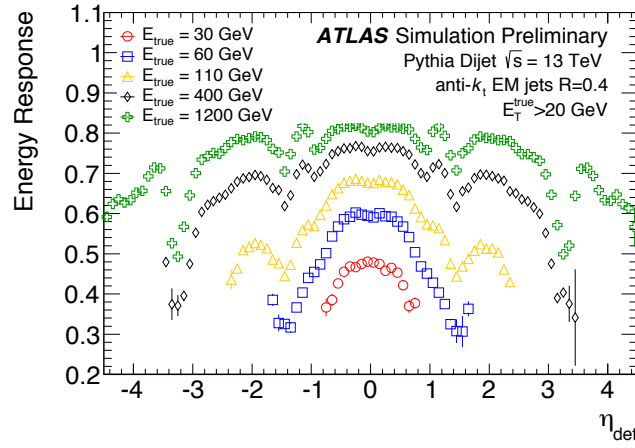


Figure 5.5: Average energy response for jets built from topoclusters at the EM scale. The response is illustrated separately for different particle-jet energies as function of the jet detector pseudo-rapidity η_{det} [123]. Points are only shown if the reconstructed jet p_T is above 7 GeV. For example, at $E_{true} = 30$ GeV, the p_T is above 7 GeV when η_{det} is below ~ 1 .

Global Sequential Calibration

The Global Sequential Calibration is a series of independent corrections that account for the residual dependence of jet energy on top of the EM scale found on longitudinal

and transverse features of the jet, mainly due to differences in the shower profiles between jets initiated by quarks and by gluons. To reduce non-Gaussian tails in the jet response distribution, a correction based on track information is applied. This uses track segments reconstructed in the muon spectrometer to identify high- p_T jets which are not fully contained in the calorimeter, referred to as punch-through. The spread of the tracks is described by the p_T weighted distance between all tracks in a jet. Considered tracks are required to have $p_T > 1$ GeV, be within the acceptance range of the ID ($|\eta| < 2.5$), and pass several basic quality criteria.

In-situ Calibration

The MC-based calibrations that are used to correct the EM scale jet may suffer from MC mismodeling. Various in-situ corrections are derived in order to cover the differences in the jet response between data and MC. They are derived from data by balancing the p_T of individual jets against well measured physics objects, and the relative difference. The in-situ corrections are derived and applied sequentially. They consist of: the η -intercalibration [126], which corrects the p_T of forward jets ($0.8 < |\eta| < 4.5$) to that of central jets ($|\eta| < 0.8$) in a dijet system up to a p_T of 1.2 TeV. The vector boson balancing [127] ($\gamma/Z + jets$), which corrects the response of central jets ($|\eta| < 0.8$) in the calorimeter to match that of well calibrated photons ($30 < p_T < 800$ GeV) and Z bosons decaying to pairs of electrons or muons ($20 < p_T < 200$ GeV). The multijet balance calibration [126], which extends the range of photon and Z boson balancing beyond the statistics driven limit of 800 GeV, where few high- p_T jets are balanced against a collection of low- p_T jets which have already benefited from the full calibration. This is performed iteratively and central high- p_T jets ($300 < p_T < 1700$ GeV) are calibrated using well calibrated lower- p_T jets.

5.3.3 Jet Energy Scale Uncertainty

Calibrations come with uncertainties which are described in the following. The full set of systematics uncertainties related to the jet energy scale are described in Ref. [128]. However, the analysis presented in this thesis considers a reduced set of 20 uncertainty

terms grouped as the following:

- Four pileup uncertainty terms to account for mismodeling in the MC simulation of the number of reconstructed primary vertices N_{PV} , the mean number of interactions per bunch crossing μ , and the energy density in jets ρ . These uncertainties are derived from both Data and MC simulations.
- Three jet-flavor related uncertainties to reflect the differences in the calorimeter response to b -quark, light-quark, and gluon-initiated jets, and the uncertainty of the jet flavor composition of the sample.
- Three uncertainty terms are associated with the η -intercalibration technique.
- One uncertainty term is derived from the single-particle response at high- p_T and applied beyond the reach of in-situ uncertainties.
- One uncertainty term associated with the punch-through correction applied in the global sequential calibration.
- Six uncertainties associated with in-situ ($\gamma/Z + jet$ techniques balance and multijet balance) are divided in different categories (statistical, detector, modeling, mixed) according to their origin.

The full combination of the uncertainties related to the jet energy scale is shown Figure 5.6. Jets with p_T of 25 GeV have a typical JES uncertainty of 5%.

5.3.4 Jet Energy Resolution

The exact energy of a jet can not be measured due to noise, stochastic fluctuations in the calorimeter response, and detector calibration effects. The jet energy measurements of the same true energy are expected to be distributed using a Gaussian spread with a width referred to as the jet energy resolution (JER). The width of the balance distributions in the η -intercalibration and vector boson in-situ calibrations are used to estimate the JER in data and MC as a function of p_T and $|\eta|$ [126], [127]. Jets with p_T of 25 GeV have a typical JER uncertainty of 3.5%.

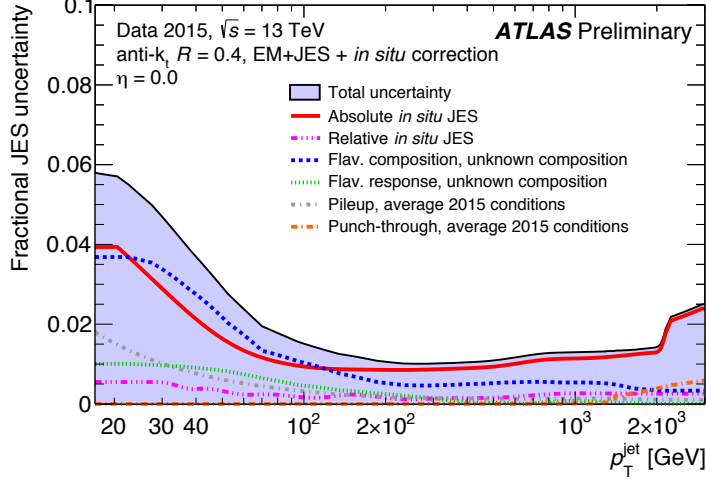


Figure 5.6: The jet energy scale uncertainty as a function of p_T at $\eta = 0$. All the uncertainty components are summed together in quadrature and the total uncertainty is shown as a filled blue region topped by a solid black line [126, 127, 129].

5.3.5 Jet Vertex Tagger

Pileup activity often creates jets that are not part of the hard scatter event and are background. Using the information from the associated tracks of a jet can help in identifying these additional jets and reducing the effect of in-time pileup. The Jet Vertex Tagger (JVT) [130] combines the information from the following two variables: corrJVF, and R_{p_T} into a multivariate analysis.

The first variable "corrJVF" is the ratio of the sum of p_T of all tracks coming from the hard scatter primary vertex (PV_0) matched to the jet. This ratio is expected to be close to one for hard scatter jets and close to zero for pileup jets since they are not originating in the PV.

The second variable R_{p_T} , is defined as the ratio of the scalar p_T sum of the tracks that are associated with the jet and originate from the hard scatter vertex, to the fully calibrated jet p_T after pileup subtraction as the following:

$$R_{p_T} = \frac{\sum_i p_T^{\text{trk}_i}(PV_0)}{p_T^{\text{jet}}}. \quad (5.6)$$

Figure 5.7 shows the distribution of the JVT discriminant output for jets originating

from the hard scatter interaction and for those from pileup. It demonstrates the separation between jets originating from the hard scatter peaking at one and jets from pileup peaking at zero.

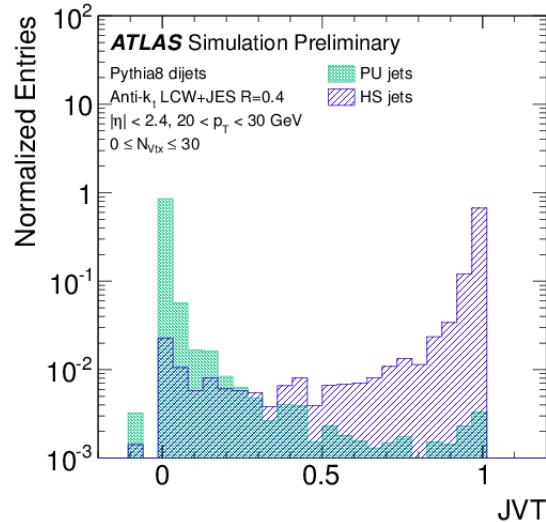


Figure 5.7: Distribution of the JVT score for hard scatter jets (the blue shaded histogram), and pileup jets (the green histogram) with $20 < p_T < 30$ GeV and $|\eta| < 2.4$ in simulated dijet events [130]. Jets with no associated tracks get a value of -0.1

In order to suppress pileup jets, a requirement of $JVT > 0.59$ is made which has a 92% selection efficiency for hard scatter jets. This requirement is only applied to jets with a p_T below 60 GeV and with $|\eta| < 2.4$ since the contribution of pileup jets at high p_T is negligible. The efficiency and the corresponding scale factors (SF) of such a cut on data and MC are derived using $Z \rightarrow \mu^+ \mu^-$ events, containing additional hard scatter jet. The associated systematic uncertainty with the JVT requirement is derived from using different MC generators to simulate $Z \rightarrow \mu^+ \mu^-$ events.

5.4 b -tagging

The identification of jets containing a b -hadron from the fragmentation of b -quarks is typically referred to as b -tagging and is of major importance for measurements with processes containing b -quarks in the partonic final states such as the analysis presented in this thesis.

b -tagging aims to identify b -jets and separate them from c - and light-jets on the basis of the longer lifetime and higher mass of the b -hadron. In the energy range above 10 GeV, the long lived b -hadrons ($\tau \sim 1.5\text{ps}$, $c\tau \sim 450\mu\text{m}$) produced in the hadronization of b -quarks can travel several millimeters, decaying at a sufficiently large distance from the production vertex to resolve a secondary vertex in the detector as shown in Figure 5.8.

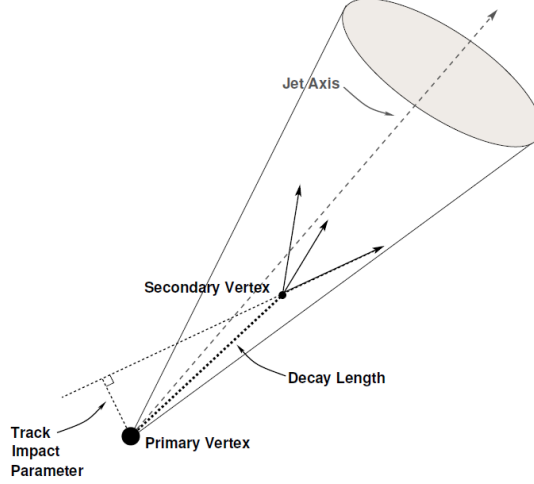


Figure 5.8: The most relevant variables (track impact parameter, primary vertex, and secondary vertex) for the identification of a jet originating from a b -quark.

Several characteristics can be used to identify this signature. The ability to identify the secondary vertex of a jet, its distance to the primary vertex⁵ (decay length) and the mass of all the associated particles to the vertex play a role in the identification of jets originating from b -quarks. Moreover, secondary vertices from the decay of b -hadrons are expected to be relatively displaced from the primary vertex and the invariant mass of the particles associated with the secondary vertex is close to 5 GeV (due to neutral decay products not being included). Instead of reconstructing the secondary vertex, the impact parameter of each track in the jet is analyzed, where the longitudinal and transverse impact parameter are defined as the minimum distance of the track to the primary vertex in the z direction and in the x - y plane, respectively. The sign of the impact parameter depends on whether the point of minimum approach to the vertex is in the same hemisphere as

5. The decay length is divided by its error to obtain the decay length significance, expressed as L/σ_L , in order to reduce the effect of poorly measured vertices.

the one defined by the jet direction or not. The impact parameter is assigned a positive sign if the track extrapolation crosses the jet direction in front of the primary vertex, and a negative one otherwise. A typical jet originating from a b -quark that has one or more tracks, and is expected to show a large and positive impact parameter significance.

5.4.1 b -tagging Algorithms

Various algorithms have been developed by ATLAS to perform the b -tagging of jets using the above described characteristics. They have been developed at multiple stages during the data taking periods, taking into consideration the various improvements of the tracking system. The output of these b -tagging algorithms are combined in a multivariate discriminant. The most relevant algorithms are described below:

- The IP3D algorithm [131] uses both the transverse and longitudinal impact parameter significances combined in a two-dimensional likelihood discriminant where their correlations are considered. Input variables are compared to templates obtained from MC simulation for both b -jet and light-jet hypotheses.
- The SV1 algorithm [131] explicitly reconstructs a displaced secondary vertex of the jet using tracks that fulfill specific quality criteria. A likelihood discriminant is constructed using various variables, such as the invariant mass of all associated tracks with the vertex, the decay length significance, the fraction of the sum of the energies of the tracks in the vertex to the sum of the energies of all tracks in the jet, and the number of two-track vertices.
- The JetFitter algorithm [132] makes full use of the topological structure of b - and c -hadron decays to reconstruct the decay chain inside the jet. It uses a Kalman-filter [133] approach in order to find a common line on which the primary vertex and vertices from the bottom or charm lie, and determines their trajectory.
- The MV2c10 algorithm [134] combines the output of the above algorithms in a Boosted Decision Tree (BDT) to achieve a better discrimination. The output of the BDT is the MV2c10 score, which is trained on b -jets as signal and a mixture of 93% light-flavor jets and 7% c -jets as background.

The performance of the b -tagging algorithms is measured by their ability to correctly identify jets coming from a real b -quark compared to the probability of mistakenly tagging a jet originating from a c -quark or a light-flavor parton (u, d, s -quark, or a gluon) as a b -jet. These quantities are usually referred to as c -tagging efficiency and mistag rate, respectively. Figure 5.9 shows the b -tagging efficiency, for the MV2c10 algorithm, with respect to the light-jet and c -jet rejection⁶.

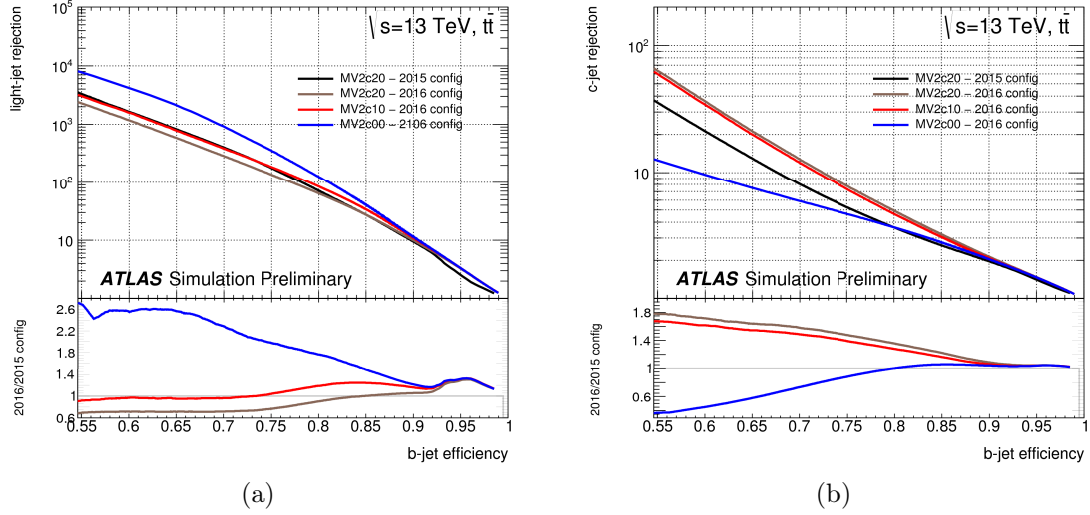


Figure 5.9: (a) Light-flavor jet and (b) c -jet rejection versus b -jet efficiency for the previous 2015 and the current 2016 configurations of the MV2 b -tagging algorithm evaluated on $t\bar{t}$ events [134]. MV2c00 stands for the MV2 algorithm where no c -jet contribution is present in the training. MV2c10 (MV2c20) denote the MV2 outputs where 7% (15%) of c -jet fractions are present in the background sample for the 2016 configuration.

The b -tagging used in this thesis relies on the MV2c10 tagger. Four operating points have been defined, based on the average b -tagging efficiency of the algorithm on simulated $t\bar{t}$ events as detailed in Table 5.2.

There are different ways of applying b -tagging. The straight forward way, which was used in previous results [135], is to choose one of the four operating points summarized in Table 5.2 with a desired b -jet efficiency and only select jets above the cut value. However, a more sophisticated approach is used here which uses the entire distribution of the MV2c10 tagger score, divided into five exclusive bins defined by the BDT values of the operating

6. The rejection factor is defined as the inverse of the efficiency to pass a given b -tagging operating point.

b -jet Efficiency [%]	BDT cut value	c -jet Rejection	Light-jet Rejection
60	0.9349	34	1538
70	0.8244	12	381
77	0.6459	6	134
85	0.1758	3.1	33

Table 5.2: Summary of the four operating points for the MV2c10 b -tagging algorithm including benchmark numbers for the efficiency and the rejection rates. The above values have been extracted from $t\bar{t}$ events with the main requirement on jet p_T to be above 20 GeV [134].

points listed in Table 5.2 and the distribution edge points interpreted as 100% and 0% efficient. This procedure is known as pseudo-continuous (PC) b -tagging. This allows for a finer differentiation of jets and division into five classes according to the purest and most efficient b -jet selection, as compared to only two classes of being tagged or not given a single operating point.

5.4.2 b -tagging Calibration

As the efficiencies of the b -tagging algorithms are derived from MC simulation, calibration is performed, correcting the efficiencies to data. The efficiency of each operating point listed in Table 5.2 has been calibrated using data samples enriched in b -, c -, and light-jets, respectively. The result of this calibration is presented in terms of scale factors $SF = \epsilon_{\text{data}}/\epsilon_{\text{MC}}$, allowing to correct for mismodeling in the input variables used in the b -tagging algorithm.

The b -jet calibration used here is derived from a high-purity sample of b -jets obtained from $t\bar{t}$ events requiring two oppositely-charged leptons in the final state. A likelihood approach [136] is used in the calibration. This achieves a precision on the b -tagging efficiency of a few percent for jet p_T between 30 and 300 GeV. Figure 5.10 shows an example of the b -jet efficiencies and SFs for the 77% operating point obtained using $t\bar{t}$ events as function of jet p_T . Most of the points illustrated in Figure 5.10 (a) are above 77% but the inclusive efficiency in the sample is 77% and most of the statistics are in the first two bins.

As mentioned above, the b -tagging algorithm also mistags c -jets and light-jets as b -jets. Therefore, the mistag efficiencies need to be calibrated as well by measuring the c -jet and

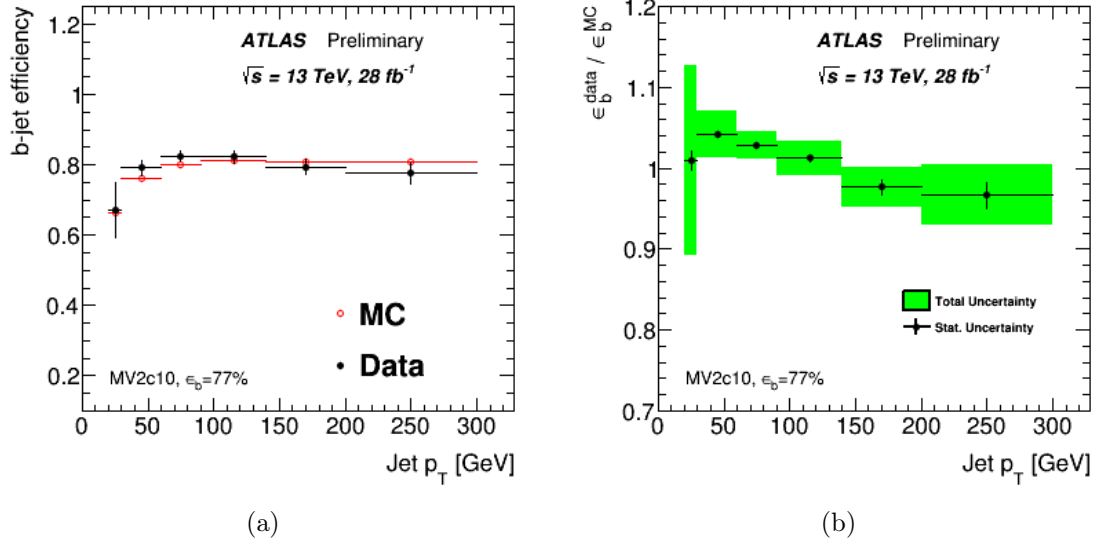


Figure 5.10: (a) b -tagging efficiency, and (b) b -tagging scale factors, ratio of the distribution in (a), for the MV2c10 algorithm with the 77% operating point as a function of jet p_T extracted from data and in simulation using $t\bar{t}$ Probability Distribution Function method [136]. Error bars indicate the combined statistical and systematic uncertainties.

light-jet tagging efficiencies and corresponding SFs to account for the differences in MC simulation and data. The tagging calibration for c -jets has been derived from the hadronic W decay in $t\bar{t}$ events [137].

The mistagging efficiency of light-jets is calculated using the negative tag method [138]. This method reverses all the internal discriminating b -tagging variables of the tagging algorithm. Then, the mistag rate can be calculated by applying the same tag weight criteria, taking into account the effects of the finite detector resolution. Due to the differences in the track resolutions in the central and more forward regions of the tracking system, the efficiency and SFs of the light-jets are calculated for two η regions separately.

5.5 Missing Transverse Energy

Momentum conservation implies that the transverse momenta of the collision products should sum to zero because the initial beam has zero transverse momentum. An inequality in the visible transverse momenta is referred to as *Missing Transverse Energy*, or E_T^{miss} . This may hint to the presence of only weakly interacting stable particles in the final state,

which traverses the detector without leaving a signal. In the case of the Standard Model these particles are known as the neutrinos.

The E_T^{miss} of an event is calculated as the magnitude of the negative vector sum of the momenta of all calibrated and reconstructed objects (hard term) and additional correction terms from the tracking (soft term) [139]. The x - and y -component of E_T^{miss} is calculated by

$$E_{x(y)}^{\text{miss}} = E_{x(y)}^{\text{miss},e} + E_{x(y)}^{\text{miss},\gamma} + E_{x(y)}^{\text{miss},\tau} + E_{x(y)}^{\text{miss},\text{jets}} + E_{x(y)}^{\text{miss},\text{SoftTerm}} + E_{x(y)}^{\text{miss},\mu}, \quad (5.7)$$

where each term is the negative sum of all the object (e, γ, τ , jet, and μ) energy projected in the x - and y -direction. The total E_T^{miss} is then given by

$$E_T^{\text{miss}} = \sqrt{(E_x^{\text{miss}})^2 + (E_y^{\text{miss}})^2}. \quad (5.8)$$

The above representation of the E_T^{miss} does not directly reflect the imbalance of the hard scattering event. For example, it does not account for the detector miscalibration, limited coverage, finite resolution, dead material and electronic noise. Furthermore, it is affected by energy deposits or tracks from pileup, cosmic rays, beam-halo or beam-gas interactions [139].

Chapter 6

ESTIMATION OF FAKE AND NON-PROMPT LEPTONS

Despite the sophisticated reconstruction algorithms, and the lepton identification and isolation requirements, described in Chapter 5, misidentification of reconstructed objects may still happen, causing background events in the analysis sample. This background is classified into: *fake leptons*, signals being selected as leptons but without a real lepton being present, and *non-prompt leptons*, real leptons that are not originating from the primary hard interaction.

This chapter presents a data-driven method based on the Matrix Method for estimating fake and non-prompt leptons. A detailed consideration of the mechanisms by which electrons and muons can be faked is described in Section 6.1. Section 6.2 briefly describes the modeling of fake events. The remainder of the chapter, Section 6.3, presents an overview of the Matrix Method to estimate the amount of fake and non-prompt leptons in the analysis sample.

6.1 Processes for Faking Electrons and Muons

Fakes and non-prompt leptons correspond to several types of reconstructed objects which satisfy the identification criteria but have different experimental signatures than leptons directly produced in the hard process.

Non-prompt leptons can occur from semi-leptonic decays of b - and c - quarks such as semi-leptonic decays of $(b \rightarrow \mu)$ or a cascade $(b \rightarrow c \rightarrow \mu)$ of B hadrons with branching ratios of $Br(b \rightarrow l^-) = (10.71 \pm 0.22)\%$, $Br(b \rightarrow c \rightarrow l^+) = (8.01 \pm 0.18)\%$ and $Br(b \rightarrow \bar{c} \rightarrow l^-) = (1.62^{+0.44}_{-0.36})\%$ [140]. These leptons are embedded in jets caused by the hadronization of the b -quark in contrast to prompt leptons, which are often produced isolated and well separated from other particles. The measurement of the detector activity around a lepton candidate (lepton isolation), defined in Section 5.2, is intended to reduce this source of background. Other sources of fake or non-prompt leptons differ between electrons and muons and are detailed in the following.

6.1.1 Sources of Fake and Non-prompt Electrons

Electrons are reconstructed depending on the presence of a reconstructed track in the inner detector matched to a deposited energy in the electromagnetic calorimeter, as described in Section 5.2.1. Misidentification of other particle types as electrons can occur and must be distinguished and suppressed. The largest backgrounds to electrons are charged hadrons from light quarks jets (u, d, s) or gluon jets. These backgrounds are separated from electrons by their hadronic shower which tends to be more diffuse than the narrow electromagnetic shower of an electron. Also, hadronic showers deposit energy in both the EM and hadronic calorimeters. While, electron's shower is typically fully contained inside the EM calorimeter.

Photon conversions into pairs $e + e^-$ that happen in the detector via interactions with material, leave both tracks and energy deposit in the electromagnetic calorimeter which are often very difficult to distinguish from prompt electrons and could be mistaken for an electron. Figure 6.1 shows the detector signature of a converted photon in the ATLAS detector leading to the misidentification of a signal electron. Also, the Dalitz decay of a high energetic π^0 mesons ($\pi^0 \rightarrow e^+e^-\gamma$), can mimic the electron signature leading to the reconstruction of a fake electron. A typical conversion background has a larger impact parameter, slightly different shower signatures, and poor track-calorimeter matching which can be used to distinguish them from prompt electrons.

6.1.2 Sources of Fake and Non-prompt Muons

Muons are reconstructed using tracks in the muon spectrometer which are matched to those in the inner detector, as described in Section 5.2.2. Muons should be the only particle type to reach the muon spectrometer. However, energetic initial charged particles with elongated shower shapes, enhances the chance of a shower particle to exit the calorimeter and enter the ATLAS muon spectrometer. Such an event is referred to as a calorimeter punch-through or particle leakage into the muon spectrometer, whose reconstructed track could be misinterpreted as a primary muon track by the muon reconstruction algorithm, causing a fake muon track. Figure 6.2 shows an initial particle and the punch-through particles of a typical calorimeter punch-through event. Non-prompt muons can occur from the in-flight disintegration of charged mesons such as charged Kaon (K^+) decaying into

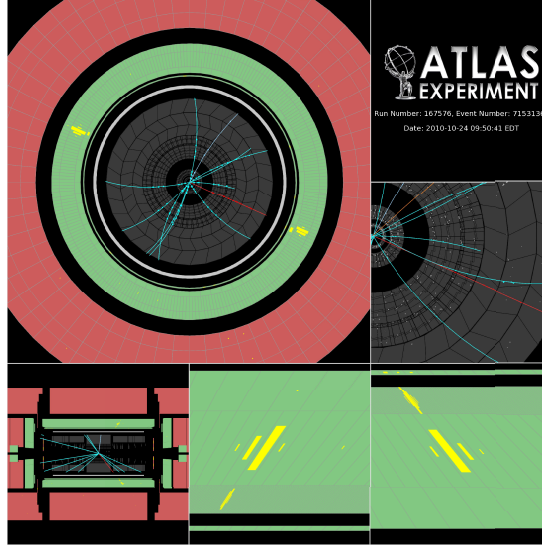


Figure 6.1: A di-photon event display selected by the $H \rightarrow \gamma\gamma$ analysis, where $m_{\gamma\gamma} = 116$ GeV. The photon conversion pair is very asymmetric, and the softer track (in red) is displayed only in the right pad [141].

$\mu\nu_\mu$.

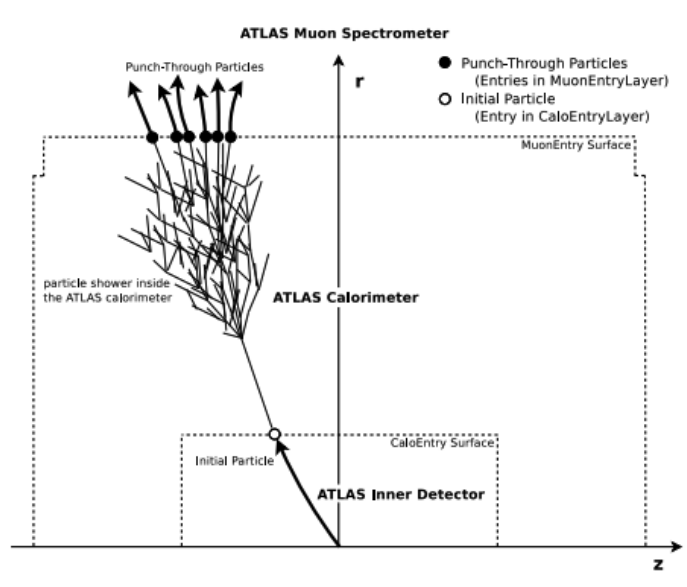


Figure 6.2: An illustration of an initial particle and the punch-through particles of a typical calorimeter punch-through event. Initial particles are depicted by empty circles and the ones arising from punch-through particles are depicted in black circles. This figure is taken from [142].

6.2 Modeling Fake Events

For simplicity, the term "fake" will be used for the sum of fake and non-prompt leptons in the remainder of this chapter. The aim here is to get an estimate of how many events are expected as a result of one or more leptons being faked, in the phase space of the analysis presented in this thesis. Assuming a tight lepton identification and isolation requirements, most fake leptons are rejected, but the small fraction of fake leptons remaining in the analysis region needs to be estimated and considered as an additional background.

Several methods exist to estimate the background arising from fake leptons. One option is to rely on the Geant4 detector simulation of the ATLAS detector and to use Monte Carlo (MC) events generated for processes expected to contribute through fake objects. However, this has two main drawbacks. Firstly, investigating a specific narrow region of a phase space would require generating a sufficiently large number of events in order to produce an estimate with a low enough statistical uncertainty which might be problematic. Secondly, background from misidentification is not expected to be accurately modeled by the MC simulation. An accurate prediction of this background would require a correct simulation of the misidentified particle and a precise model of the rate of misidentification, keeping in mind that only a very small fraction of jets fake leptons. In order to model this rate correctly, an accurate modeling of the non-Gaussian tails of the detector response to jets is required. However, this level of details is not expected from the MC simulation. For these reasons, data-driven methods are favored.

The method used in this thesis is based on the Vanilla Matrix Method, which was originally developed in the *D0* experiment at Tevatron in 2007 [143]. It is one of the most used estimates for many analyses at the ATLAS experiment [144–149]. The Matrix method developed in the context of measurements related to top-quark production, is detailed in the following section.

6.3 The Vanilla Matrix Method

The Matrix Method utilizes two populations of events with leptons passing tight and loose identification criteria for the event selection as illustrated in Figure 6.3.

Events with leptons passing the analysis selections are referred to as the "Tight" sample

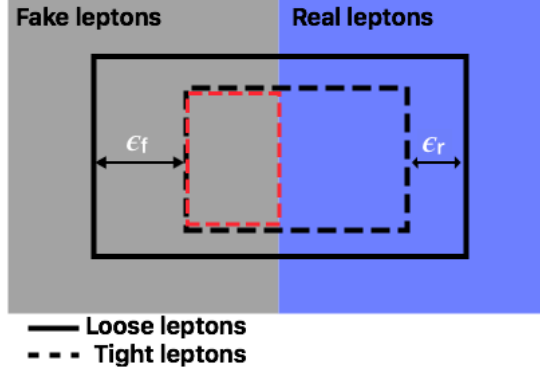


Figure 6.3: Illustration of the Matrix Method: the total event sample consists of events with Loose (solid box) and Tight (dashed box) leptons. The red dashed box represent the analysis region where fake leptons could occur and need to be estimated. The efficiency ϵ_r and ϵ_f are determined experimentally using regions enriched in real and fake leptons, respectively.

(the dashed box in Figure 6.3). By loosening those selection requirements such as lepton identification and isolation, a "Loose" sample is defined (the solid box in 6.3). Hence the Tight sample is a subset of the Loose sample. As illustrated in Figure 6.3 the Loose and Tight samples consist of both real (r), and fake (f) leptons. Therefore the number of events with a loose lepton (N^l) and the number of events with a tight lepton (N^t) can be expressed as a linear combination of both the real (N_r) and fake (N_f) leptons, as the following:

$$N^l = N_r^l + N_f^l \quad (6.1)$$

$$N^t = N_r^t + N_f^t \quad (6.2)$$

The fraction of real leptons in the loose selection which also passes the tight requirements is defined as the real efficiency

$$\epsilon_r = \frac{N_r^t}{N_r^l} \quad (6.3)$$

and similarly, the fake efficiency

$$\epsilon_f = \frac{N_f^t}{N_f^l} \quad (6.4)$$

refers to the fraction of loose fake leptons that passes the tight requirements. Since the Tight sample is a subset of the Loose sample, ϵ_r and ϵ_f are by definition $\in [0; 1]$. Then equation 6.2 can be expressed as:

$$N^t = \epsilon_r N_r^l + \epsilon_f N_f^l \quad (6.5)$$

The number of fake background events in the final analysis selection (the red dashed box in Figure 6.3) can be measured from combining equation 6.2 and 6.5 into the following expression:

$$N_f^t = \frac{\epsilon_f}{\epsilon_r - \epsilon_f} (\epsilon_r N^l - N^t). \quad (6.6)$$

Therefore, the number of fake leptons can be estimated if the real ϵ_r and fake ϵ_f efficiencies are known and the number of tight and loose events are counted in the data sample.

Equation 6.6 gives an integrated yield of the fake background, whereas, the measurement is performed in binned distributions. Therefore, Equation 6.6 can be expressed as a weighting factor which will be applied to each data event in the Loose sample to estimate the distribution of the fake background in the analysis. The weighting factor is expressed as the following:

$$w_i = \frac{\epsilon_f}{\epsilon_r - \epsilon_f} (\epsilon_r - \delta_i), \quad (6.7)$$

where (i) stands for the event, δ_i equals unity if the loose event (i) passes also the tight requirement and zero otherwise.

Equation 6.6 can then be expressed as the following

$$N_f^t = \sum_i w_i N^l. \quad (6.8)$$

If a loose lepton passes also the tight selection, the event weight in Equation 6.7 will be negative. Tight leptons are highly expected to be real leptons which should be subtracted from the background estimate, i.e. they contribute to the negative event weights of the Matrix Method. While, loose only leptons are more likely to be fake leptons, i.e. background events contributing to the positive weights of the Matrix Method. The larger the difference between the Loose and the Tight sample (the difference between the loose (solid box) and the tight (dashed box) in Figure 6.3), the less likely it is to end up with large amount of negative weights.

The choice of the Loose sample is essential for the performance of the Matrix Method. The definition of the Loose sample should contain the Tight sample and should account properly for the possible sources of the extra leptons. This is needed in order to not bias the measurement of the fake (ϵ_f) efficiency to a particular source.

In ideal cases, both the real (ϵ_r) and fake (ϵ_f) efficiencies should be measured in the analysis regions. However, the amount of fake events is largely suppressed due to the efficient background suppression of the analysis selection and is hard to be distinguished from real leptons. Therefore, the fake ϵ_f (real ϵ_r) efficiencies are measured in data using dedicated regions which are enriched in real and fake leptons, respectively. The following section details the measurement of fake and real efficiencies.

6.3.1 Fake and Real Efficiencies

The fake efficiency ϵ_f is measured in data samples enriched with fake lepton events, referred to as the fake-enriched regions CR_f . Fake-enriched regions are chosen as close kinematically as possible to the signal regions in order to ensure that the efficiencies derived are applicable in the analysis regions. Fake-enriched regions are picked with a set of leptons that are almost surely fake, and the fake efficiency is calculated as the ratio of events with tight leptons over events with loose leptons. However, there will be a contamination from real leptons which is estimated using MC simulations. The MC simulation contain all relevant SM processes such as $t\bar{t}$, single-top, W/Z +jets, and dibosons. Then, the fake efficiencies are measured by taking the ratio in yields between tight and loose events after

subtracting the sum of simulated background from data, as the following:

$$\epsilon_f = \frac{N_f^t}{N_f^l} = \frac{N_{\text{data}}^t - N_{\text{MC}}^t}{N_{\text{data}}^l - N_{\text{MC}}^l} \quad (6.9)$$

Fake efficiencies are measured separately for electrons and muons, using different fake-enriched regions for each. In the single lepton channel, events with W decays would have prompt leptons and neutrinos. Therefore, missing transverse momentum (E_T^{miss}) and transverse mass of the lepton (m_T^W) are good discriminating variables, given that fake events tend to have low E_T^{miss} and m_T^W . The transverse mass is defined as

$$m_T^W = \sqrt{2p_T E_T^{\text{miss}}(1 - \cos\Delta\phi)}, \quad (6.10)$$

where $\Delta\phi$ is the difference in azimuthal angle between the lepton and E_T^{miss} . In the case of electrons, fake-enriched regions are defined requiring low E_T^{miss} and/or low m_T^W .

In the case of muons, fakes are mostly expected from the decay of a b -hadron within a jet, which is produced at significant displacement from the primary vertex. The impact parameter of the muon with respect to the primary vertex is expected to be larger than the prompt muons. Therefore, a good discriminant to probe fake muons from b -hadron decays is the muon impact parameter significance $d_0^{\text{sig}} = \frac{d_0}{\sigma_{d_0}}$.

The real efficiencies ϵ_r are measured using the tag-and-probe method implemented on data for the $Z \rightarrow \mu\mu$ and $Z \rightarrow ee$ selection where a pure lepton sample can easily be selected, as described in Section 5.2. The tags are the leptons passing the tight selection requirements. While, the probes are the leptons passing the loose selection requirements. The number of all probes (events with a tight and a loose lepton) is the denominator of the efficiency and the number of probes which pass the tight criteria is the numerator.

Assuming that the efficiencies vary only as a function of the kinematic properties of the object, it is possible to determine the fake background in the fake-enriched regions and extrapolate them to the signal regions. Therefore, the real ϵ_r and fake ϵ_f efficiencies can be parametrized as function of the kinematic properties of the event such as lepton p_T , lepton η , and leading jet p_T .

Then, the different combinations of the variables are parametrized through:

$$\epsilon_k(x_1, \dots, x_N; y_1, \dots, y_M) = \frac{1}{\epsilon_k(x_1, \dots, x_N)^{M-1}} \cdot \prod_{j=1}^M \epsilon_k(x_1, \dots, x_N; y_j), \quad (6.11)$$

where k represents the real and fake efficiencies, and the number of x and y variables is represented by N and M , respectively.

The expression $\epsilon_k(x_1, \dots, x_N)$ represents the efficiency measured as a function of the x variables. While the $\epsilon_k(x_1, \dots, x_N; y_j)$ represents the efficiency measured as a function of the x variables and of the variable y_j . Equation 6.11 entails that the full correlation among the variables x (typically discrete variables such as number of jets) and each of the variables y (typically continuous variables like p_T , and η with relatively large number of bins) is taken into account. For each of the real or fake efficiency ϵ_k , only a sub-set of the y variables is used. Typical y variables are the lepton p_T and η , the p_T of the leading jet in the event ($p_T^{\text{leading jet}}$), the angular distance between the lepton and the missing energy in the event ($\Delta R(l, \text{jet})$), and the angular distance in the transverse plane between the lepton and the missing energy in the event ($\Delta\phi(l, E_T^{\text{miss}})$). At the end, the choice is driven by the observed dependencies, and the stability of the fake lepton estimate.

Chapter 7

SEARCH FOR THE PRODUCTION OF A STANDARD MODEL HIGGS BOSON IN ASSOCIATION WITH TOP-QUARKS AND DECAYING INTO A PAIR OF BOTTOM-QUARKS

The production of the Higgs boson in association with top-quarks, $t\bar{t}H$, provides a distinctive access to the Yukawa coupling of the Higgs boson to the top-quark. The measurement of this coupling is essential to assess the SM behaviour of the observed Higgs boson.

This chapter describes the search for the $t\bar{t}H$ production where the Higgs boson decays into $b\bar{b}$. This search uses the data collected by the ATLAS detector during 2015 and 2016 at a center-of-mass energy of 13 TeV. The analysis has been published in [150].

An overview of the advantages and the challenges of the Higgs boson decaying into $b\bar{b}$ channel are discussed in Section 7.1. A review of the ATLAS and CMS analyses at 8 TeV in this channel is in Section 7.2. The used data and simulated samples are explained in Section 7.3. The object selection is described in Section 7.4 and the event selection is explained in Section 7.5. The general analysis strategy to separate signal and background events using Boosted Decision Tree (BDT) techniques, based on various kinematic variables, is discussed in Section 7.6. The estimation of the background is introduced in Section 7.7. Kinematic distributions in the various analysis regions are presented in Section 7.8. The systematic uncertainties of the measurement are presented in Section 7.9. The search for the $t\bar{t}H(H \rightarrow b\bar{b})$ production is expressed in terms of the signal strength parameter μ , which is defined as the ratio of the observed to the expected number of signal events assuming the SM cross section. The signal strength is extracted from a likelihood fit performed simultaneously in all the analysis regions, as described Section 7.10.

7.1 Measurement of $t\bar{t}H$ in the $(H \rightarrow b\bar{b})$ Decay Mode

The production of the Higgs boson in association with a $t\bar{t}$ pair contributes to only about 1% of the total Higgs boson production cross-section at the LHC. It has a cross-section of $0.507 \text{ pb }^{+5.8}_{-9.2}$ (QCD scale) ± 3.6 (PDF + α_s) at 13 TeV [18]. However, the measurement of

$t\bar{t}H$ represents an essential ingredient in the measurement of the Yukawa coupling to the top-quark. This production mode will allow to probe directly the top Yukawa coupling from a tree-level diagram.

The number of $t\bar{t}H$ events ($N_{t\bar{t}H \text{ events}}$) are proportional to the luminosity (L), the $t\bar{t}H$ cross section ($\sigma_{t\bar{t}H}$), and the branching ratio of the Higgs boson ($B(H)$) and the top-quark pair ($B(t\bar{t})$) decay modes, as shown in the following equation:

$$N_{t\bar{t}H \text{ events}} = L \cdot \sigma_{t\bar{t}H} \cdot B(H) \cdot B(t\bar{t}) \cdot \epsilon \cdot A, \quad (7.1)$$

where ϵ is the selection efficiency and A is the acceptance. Given the small production cross section and the available integrated luminosity, decay modes with the largest branching ratio are the most promising ones. For a SM Higgs boson mass of 125 GeV, the decay into a pair of b -quarks has the largest branching fraction of about 58% [20].

Events in the $t\bar{t}H(H \rightarrow b\bar{b})$ analysis are split into three different channels based on the decay of the top-quark pair. In the Standard Model, the top-quark decays almost 100% of the cases into a W boson and a b -quark [20], where the W boson further decays leptonically or hadronically. The single-lepton¹ channel, where one W boson decays leptonically into e or μ with their corresponding neutrinos ν_e, ν_μ and the other W boson decays hadronically into two quarks, corresponds to about 30% of the branching ratio. The dilepton channel, where both W bosons decay leptonically, corresponds to about 4%. The full hadronic channel, where both W bosons decay hadronically, corresponds to the largest branching ratio of about 46%. These branching ratios exclude the hadronic and leptonic decaying τ (for more details see Section 2.5). The analysis presented in this thesis considers the single-lepton and dilepton channels. Despite the lower statistics, the single-lepton and dilepton channels are preferred over the full hadronic channel due to the significant lower background arising from multijet processes, and the ability to trigger on events with at least one lepton.

The $H \rightarrow b\bar{b}$ decay mode has several challenges which are common to the single-lepton and dilepton channels. The biggest challenge lies in the attempt to reconstruct a complex signature, with several jets and b -tagged jets in the final state, over a large background arising from the production of $t\bar{t} + b\bar{b}$ process. Figure 7.1 shows two Feynman diagrams of

1. In this analysis, the term "lepton" refers to electrons or muons.

$t\bar{t}H$ production, in which the Higgs boson is either formed by top-quark fusion (Figure 7.1 (a)) or is radiated off a top-quark (Figure 7.1 (b)). Figure 7.2 shows the Feynman diagram for the dominant $t\bar{t} + b\bar{b}$ background that is similar in the kinematics to the $t\bar{t}H(H \rightarrow b\bar{b})$ signal, and has a cross-section approximately one or two orders of magnitude larger than $t\bar{t}H(H \rightarrow b\bar{b})$ depending on the analysis phase space [151].

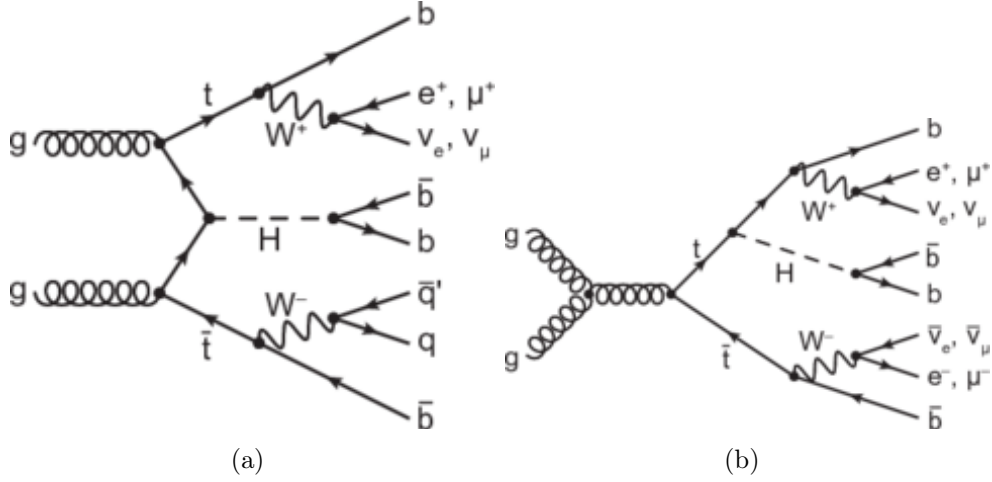


Figure 7.1: Feynman diagram representation at tree-level for (a) t-channel and (b) s-channel of the Higgs boson in association with a top-quark pair ($t\bar{t}H$) and the subsequent decay of the Higgs boson to $b\bar{b}$. Moreover, (a) represents the single-lepton, and (b) represents the dilepton final state configuration of the $t\bar{t}H(H \rightarrow b\bar{b})$ channel.

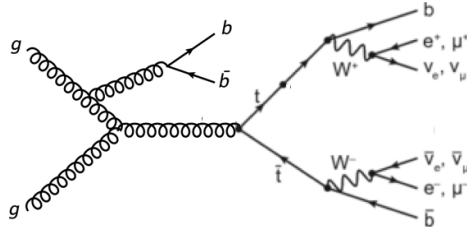


Figure 7.2: Feynman diagram representation of the main background $t\bar{t} + b\bar{b}$.

7.2 Search for $t\bar{t}H(H \rightarrow b\bar{b})$ at 8 TeV

ATLAS searched for $t\bar{t}H(H \rightarrow b\bar{b})$ at $\sqrt{s} = 8$ TeV, using $t\bar{t}$ decays with one or two electrons or muons [152] or zero leptons (the full hadronic channel) [153]. The individual

measurements are consistent with each other and the measured signal strengths (μ) are compatible with the SM expectations. A combined signal strength μ of 1.4 ± 1.0 was measured [153]. The central value and uncertainty on the signal strength is driven by the single-lepton analysis and the largest systematic effect arises from the uncertainty in the normalization of the $t\bar{t} + b\bar{b}$ background.

The CMS collaboration has searched for the same process at $\sqrt{s} = 7$ TeV, and $\sqrt{s} = 8$ TeV using the single-lepton and dilepton $t\bar{t}$ decay modes, obtaining a signal strength μ of 0.7 ± 1.9 [154].

7.3 Data and Simulation Samples

7.3.1 Data Taking

This analysis uses a set of data events collected by the ATLAS detector in pp collisions at the LHC in 2015 and 2016 at a center-of-mass energy of 13 TeV. The time evolution of the total integrated luminosity delivered to and recorded by ATLAS during stable beams for pp collisions at 13 TeV center-of-mass energy in 2016 is shown in Figure 7.3 (a). The delivered luminosity (green in Figure 7.3 (a)) stands for luminosity delivered from the start of stable beams until the LHC requests ATLAS to change the settings of the detector to a safe standby mode to allow for a beam dump or beam studies. The recorded luminosity (yellow in Figure 7.3 (a)) reflects the Data Acquisition (DAQ) inefficiency and the inefficiency of the so-called "warm start"². The ATLAS data acquisition efficiency is 92.1% in 2015 and 92.4% in 2016. The dataset corresponds to an integrated luminosity of $3.2 \pm 0.1 \text{ fb}^{-1}$ recorded in 2015, and $32.9 \pm 0.7 \text{ fb}^{-1}$, recorded in 2016, for a total of $36.1 \pm 0.8 \text{ fb}^{-1}$ [155].

The mean number of interactions per crossing corresponds to the mean of the Poisson distribution of the number of interactions per crossing calculated for each bunch. It is calculated from the instantaneous per bunch luminosity as $\langle \mu \rangle = L_{\text{bunch}} \times \sigma_{\text{inel}} / f_r$ where L_{bunch} is the measured instantaneous luminosity per number of colliding bunch pairs, σ_{inel} is the inelastic cross section for pp interactions which is 80 mb for 13 TeV

2. Warm start refers to the time from when stable beam is declared until when the tracking detectors ramp of the high-voltage for the pixel system to turn on its preamplifiers.

collisions [86], and f_r is the LHC revolution frequency. In 2015, the average number of interactions per bunch is measured to be $\langle \mu \rangle = 13.7$, which increased to $\langle \mu \rangle = 24.9$ in the 2016 data taking period. Figure 7.3 (b) shows the bunch crossing $\langle \mu \rangle$ for the combined 13 TeV data from 2015 and 2016. The increased number of interactions per bunch crossing also results in a higher number of energy deposits in the detector which are not originating from the hard scattering process of interest. This pileup can influence the object reconstruction, if not identified and treated accordingly.

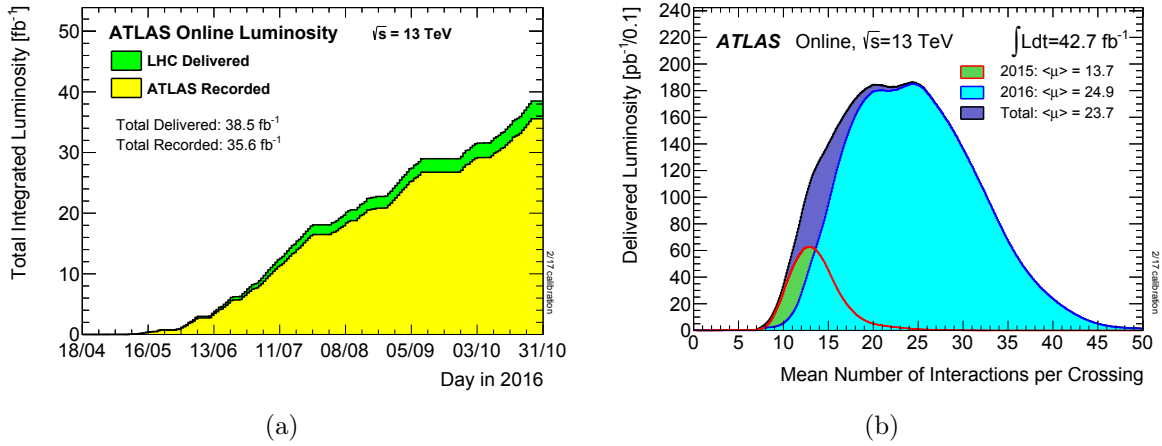


Figure 7.3: (a) Cumulative luminosity versus time delivered to (green), recorded by ATLAS (yellow), during stable beams in pp collisions at 13 TeV center-of-mass energy in 2016. (b) The luminosity-weighted distribution of the mean number of interactions per crossing for the 2015 and 2016 pp collision data. All the data delivered to ATLAS during stable beams, the integrated luminosity and mean $\langle \mu \rangle$ value are shown (see Section 7.3 for more details). The plots are taken from [86].

The dataset is separated into periods according to the running conditions such as beam settings and trigger configurations. Only periods in which all the sub-detectors are fully functional, referred to as the Good Run List (GRL), are considered for this analysis. About 17.9% (7.5%) of the events in 2015 (2016) do not satisfy the GRL.

7.3.2 Triggers

This analysis is based on events where the detector read-out is triggered by the presence of at least one electron or one muon [156], referred to as single-lepton triggers, with p_T above 24 GeV (26 GeV) for the 2015 (2016) data taking. Single-lepton triggers are chosen since at

least one electron or one muon are expected from the single or dileptonic top-quark decay. Table 7.1 summarizes the triggers used in the analysis for the 2015 and 2016 data taking periods. The 2016 triggers with the lower- p_T threshold include isolation requirements on the candidate lepton. This isolation requirement is applied in order to keep the trigger rate under control and to reduce the high trigger rate of leptons produced in hadron decays. Isolation requirements also reduce the amount of fake and non-prompt lepton background. At high p_T threshold, this background is not significant, therefore the isolation requirement can be dropped to increase the trigger efficiency. Events are required to pass the logical OR of the triggers listed in Table 7.1.

Event Filter Menu	Online Object	p_T [GeV]
2015		
e24_lhmedium _ L1EM20VH	electron	24
e60_lhmedium	electron	60
e120_lhloose	electron	120
mu_20_iloose_L1MU15	muon	20
mu_50	muon	50
2016		
e26_lhtight_nod0_ivarloose	electron	26
e60_lhmedium_nod0	electron	60
e140_lhloose_nod0	electron	140
mu_26_ivarmedium	muon	26
mu_50	muon	50

Table 7.1: Single electron and muon triggers used in the analysis for the 2015 and 2016 data taking. "Online" refers to the object used in the trigger logic. The electron identification operating points are represented by "lhtight", "lhmedium", "lhloose", and "loose" and the isolation operating points are represented by "ivarloose" and "ivarmedium" (see details in Section 5.2). "nod0" refers to absence of the track impact parameter requirement. "L1EM20VH" stands for the seed of lowest un-prescaled single electron trigger where "V" refers to the η -dependent threshold, "H" refers to the hadronic isolation. Similarly, "L1MU15" stands for the seed of lowest un-prescaled single muon trigger.

7.3.3 Simulated Samples

Simulated event samples obtained with Monte Carlo (MC) event generators are used in this analysis to estimate the signal and background contributions, to calculate detector acceptance, and to train the Boosted Decision Trees (BDTs).

Signal Samples

The $t\bar{t}H$ signal process is modeled using MADGRAPH5_AMC@NLO [67] (referred to in the following as MG5_aMC@NLO) version 2.3.2 with a NLO matrix element (ME), interfaced to the PYTHIA 8.210 [69] parton shower (PS) model using the A14 [157] set of tunable parameters. This sample is produced inclusive in Higgs boson decays with the NNPDF3.0NLO [158] parton distribution function (PDF) set using factorization and renormalization scales set to $\mu_F = \mu_R = H_T/2$, where H_T is defined as the scalar sum of the transverse masses $\sqrt{p_T^2 + m^2}$ of all final state particles. The Higgs boson mass is set to 125 GeV, and the top-quark mass is set to $m_t = 172.5$ GeV. The top-quarks are decayed using MADSPIN [159] and preserve all spin correlations. The $t\bar{t}H$ cross-section and the Higgs boson decay branching fractions are taken from (N)NLO theoretical calculations [18, 160–164].

Samples for $t\bar{t}$ + jets Background

For the modeling of the $t\bar{t}$ +jets process, several generators were used, with different perturbative accuracy and covering a range of choices for the parton shower, hadronization, PDF, and underlying event tune. This is done in order to study potential biases due to particular model components.

The nominal $t\bar{t}$ +jets sample is generated using the POWHEG-BOX v2 NLO generator [65, 165–168] with the NNPDF3.0 PDF set. The h_{damp} parameter that controls the p_T of the first additional emission beyond the Born configuration, is set to 1.5 times the top-quark mass. The parton shower and hadronization are modeled by PYTHIA 8.2 with the appropriate A14 [157] set of tunable parameters. The normalization and factorization scales are set to the transverse mass of the top-quark, defined as $m_{T,t} = \sqrt{m_t^2 + p_{T,t}^2}$, where $p_{T,t}$ is the transverse momentum of the top-quark in the $t\bar{t}$ center-of-mass reference frame. The sample is normalized using the Top++ 2.0 [39] inclusive cross section of 832^{+46}_{-51} pb, obtained from next-to-next-to-leading order (NNLO) in QCD including resummation of next-to-next-to-leading logarithmic (NNLL) soft gluon terms [34, 36–38].

The impact of variations of the amount of additional radiation, is assessed using two additional $t\bar{t}$ +jets samples generated using different settings for POWHEG and variations of the PYTHIA 8 Var3c A14 tune variations [169]. The A14 tune variations correspond to the varying of α_s that impacts ISR in the A14 tune. The samples are generated with the following setup:

- The sample with reduced QCD radiation is generated where the factorization and renormalization scales are multiplied by a factor of 2.0, the h_{damp} value stays at $1.5 m_t$ and the Var3c down variation from the A14 tune is used.
- The sample with increased QCD radiation is generated where the factorization and renormalization scales are multiplied by a factor of 0.5, the h_{damp} value is increased to $3.0 m_t$ and the Var3c up variation from the A14 tune is used.

To assess the effect of modeling of the parton shower, hadronization, and underlying event on the measurement, the aforementioned $t\bar{t}$ +jets sample is also interfaced with HERWIG 7 [71] version 7.0.1, with H7-UE-MMHT set of tunes parameters for the underlying event.

Predictions of the above MC generators are compared to ATLAS data of inclusive top-quark pair production, in which unfolded distributions from 8 and 13 TeV measurements are taken into account. Figure 7.4 shows an example of the POWHEG+PYTHIA 8 samples with different tune variations compared to data at $\sqrt{s} = 13$ TeV. The MC simulation setup of the top-quark pair production has been studied for the modeling of the POWHEG generator interfaced to the PYTHIA 8 and HERWIG 7 shower generators [170, 171]. Studies using unfolded data from the ATLAS analyses at $\sqrt{s} = 8$ TeV and $\sqrt{s} = 13$ TeV showed that POWHEG+PYTHIA 8 is the MC generator that models $t\bar{t}$ production very well [170, 171].

Furthermore, an alternative $t\bar{t}$ +jets sample is generated with SHERPA using matrix element with multiple partons. This sample is used to simultaneously assess the NLO generator, the number of partons in the matrix element calculation, the parton shower, the hadronization model, and the underlying event. This alternative $t\bar{t}$ +jets sample is generated using SHERPA version 2.2.1 with ME+PS@NLO setup, interface with OPENLOOPS, providing a matrix element calculation with NLO accuracy up to one additional parton and LO accuracy up to four additional partons. The NNPDF3.0NNLO PDF set was used and both the renormalization and factorization scales were set to $\sqrt{(0.5 \times (m_{T,t}^2 + m_{T,\bar{t}}^2))}$. This sample employs the 5 flavor scheme where additional b -quarks are considered massless

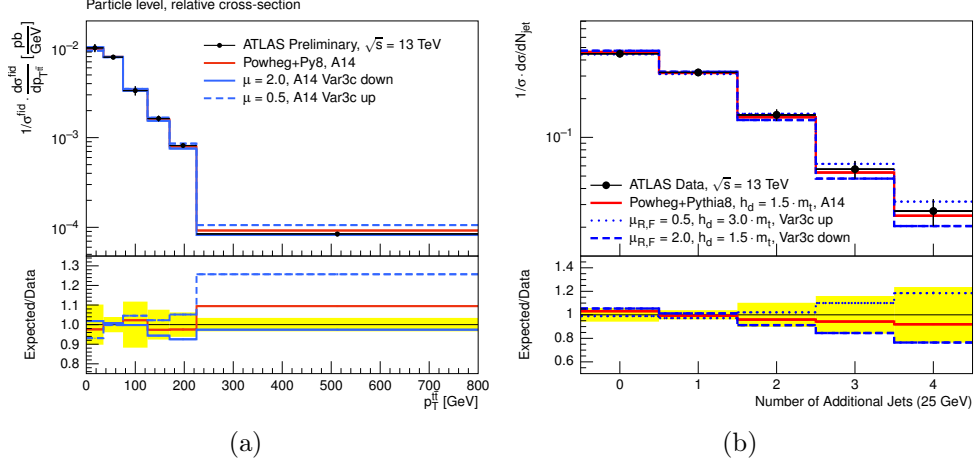


Figure 7.4: The POWHEG+PYTHIA 8 samples with different h_{damp} variations are compared to data at $\sqrt{s} = 13$ TeV. The comparison is performed for (a) the transverse momentum of the $t\bar{t}$ system and (b) for the number of additional jets, using ATLAS data unfolded to particle level from the analysis published in [172]. These figures are taken from [170, 171].

in the calculation of the matrix element, and is referred to as SHERPA5F in the remainder of this thesis.

A sample offering a description of the $pp \rightarrow t\bar{t}b\bar{b}$ process in terms of the matrix elements, is also used. NLO predictions with massive b -quarks in the four-flavor number scheme (4FNS) matched to a parton shower [173] are available in the SHERPA+OPENLOOPS [73, 75], referred to as SHERPA4F. The SHERPA4F sample uses SHERPA version 2.1 and the CT10 [66, 174] 4FNS PDF set. The renormalization scale is set to the CMMPS [173] value, $\mu_{\text{CMMPS}} = \prod_{i=t,\bar{t},b,\bar{b}} E_{T,i}^{1/4}$, and the factorization scale is set to $H_T/2 = \frac{1}{2} \sum_{i \in FS} E_{T,i}$. The resummation scale μ_Q , which sets an upper bound for the hardness of the parton shower emissions, is set to $H_T/2$.

The top-quark mass in the $t\bar{t}$ -jets samples is set to $m_t = 172.5$ GeV. Table 7.2 contains a list of the settings used for the simulation $t\bar{t}$ samples that are used in this analysis. All of these samples are normalized to the inclusive $t\bar{t}$ cross-section calculated at NNLO+NNLL accuracy [34, 36–38].

Samples for Other Backgrounds

Backgrounds arising from W/Z -jets events, and diboson production in association

ME gen. PS/UE gen.	POWHEG-BOX PYTHIA 8	POWHEG-BOX PYTHIA 8	POWHEG-BOX PYTHIA 8	POWHEG-BOX HERWIG 7	SHERPA5F	SHERPA4F
Ren. scale	$m_{T,t}$	$\frac{1}{2} \cdot m_{T,t}$	$2 \cdot m_{T,t}$	$m_{T,t}$	$\sqrt{\frac{m_{T,t}^2 + m_{T,\bar{t}}^2}{2}}$	μ_{CMMPS}
Fact. scale	$m_{T,t}$	$\frac{1}{2} \cdot m_{T,t}$	$2 \cdot m_{T,t}$	$m_{T,t}$	$\sqrt{\frac{m_{T,t}^2 + m_{T,\bar{t}}^2}{2}}$	$H_T/2$
h_{damp}	$1.5 \cdot m_t$	$3 \cdot m_t$	$1.5 \cdot m_t$	$1.5 \cdot m_t$	—	—
ME PDF Tune	NNPDF3.0NLO A14	NNPDF3.0NLO A14 Var3c up	NNPDF3.0NLO A14 Var3c down	NNPDF3.0NLO H7-UE-MMHT	NNPDF3.0NNLO Default tune	CT10 4F Default tune

Table 7.2: Summary of the settings used for the simulation of the $t\bar{t}$ +jets samples. For the renormalization and factorization scales, $m_{T,t} = \sqrt{m_t^2 + p_{T,t}^2}$ ($m_{T,\bar{t}} = \sqrt{m_t^2 + p_{T,\bar{t}}^2}$) indicates the transverse mass of the top (anti-top) quark, where $p_{T,t}$ ($p_{T,\bar{t}}$) is the transverse momentum of the top (anti-top) quark in the $t\bar{t}$ center-of-mass reference frame. The SHERPA4F $t\bar{t} + b\bar{b}$ sample, in the last column $\mu_{\text{CMMPS}} = \prod_{i=t,\bar{t},b,\bar{b}} E_{T,i}^{1/4}$ and $H_T/2 = 1/2 \sum_{i=t,\bar{t},b,\bar{b}} E_{T,i}$.

with jets, are generated using the SHERPA 2.2.1 generator. In the W/Z +jets samples, the matrix elements are calculated up to two partons at NLO and four partons at LO using the Comix [175] and OPENLOOPS matrix element generators and merged with SHERPA parton shower [74] using the ME+PS@NLO prescription [176]. The CT10 PDF set is used. The W/Z +jets cross sections are normalized to the NNLO calculations [177]. The diboson+jets samples are generated using the same approach but with up to one (ZZ) or zero (WW, WZ) additional partons are NLO and up to three additional partons at LO. They are also normalized to their respective NLO cross sections.

Single top, Wt and s -channel background samples are generated using POWHEG-BOX v1 at NLO accuracy using the CT10 PDF set. Some of the Feynman graphs that contribute to the Wt channel can be interpreted as the production of a $t\bar{t}$ pair production at LO, with subsequent decay of the \bar{t} into a $\bar{b}W$ pair. In order to avoid this, the overlap between the $t\bar{t}$ and the Wt final states is removed using the "diagram removal" scheme [178]. The electroweak t -channel single top events are generated using the POWHEG-BOX v1 generator which uses the four-flavor scheme for the NLO matrix elements calculations together with the fixed four-flavor PDF set CT10 4F. In this process, the top-quarks are decayed using MadSpin [159] which preserves all the spin correlations. All the single top samples are interfaced to PYTHIA 6.428 with Perugia 2012 underlying-event tune. The single top, Wt , t - and s -channel samples are normalized to the approximate NNLO theoretical cross sections [179–181].

Sample	ME Generator	PDF	Parton Shower Generator
$t\bar{t}H$	MG5_aMC@NLO	NNPDF3.0NLO	PYTHIA 8.2
$t\bar{t} + \text{jets}$	POWHEG	CTEQ6L1	PYTHIA 8.2
$W + \text{jets}$	SHERPA	CT10	SHERPA 2.2.1
$Z + \text{jets}$	SHERPA	CT10	SHERPA 2.2.1
Single top (s-channel, Wt)	POWHEG	CT10	PYTHIA 6.428
Single top (t-channel)	POWHEG	CT10f4	PYTHIA 6.428
$t\bar{t}V$	MG5_aMC@NLO	NNPDF3.0NLO	PYTHIA 8.2
Diboson	SHERPA	CT10	SHERPA 2.1.1

Table 7.3: The MC generators and the parameters used to simulate processes considered in this analysis as signal and backgrounds, see text for further details.

Samples for $t\bar{t}V$ ($t\bar{t}W, t\bar{t}Z$) events are generated at NLO in the matrix-element calculations using MG5_aMC@NLO interfaced with PYTHIA 8.210 with NNPDF3.0NLO PDF and the A14 tune.

The event generators used for the signal and background samples are listed in Table 7.3, together with the used PDF settings.

Common Settings

Decays of b - and c -hadrons in the above described samples are generated using EVTGEN v1.2.10 [72], except samples which are simulated by the SHERPA generator. EVTGEN handles decays of b - and c -hadrons taking into account up-to-date information about decay modes and branching fractions.

The event reconstruction is affected by multiple pp collisions called pileup. In order to simulate the effects of pileup, additional soft QCD interactions are generated using PYTHIA 8.186 [69] with the A14 tune and overlaid onto the simulated hard scatter event. All the MC samples are simulated taking into account the pileup conditions in the 2015 and 2016 data. Therefore, simulated events are re-weighted so that the distribution of the average number of pp interactions per bunch crossing matches that observed in data.

The generated particles of most of the MC samples are propagated through the full ATLAS detector simulation [182] based on Geant4 [76], as discussed in Section 3.8. A faster simulation, where the full Geant4 simulation of the calorimeter response is replaced by a detailed parametrization of the calorimeter shower shapes [183], is adopted for some of the MC samples used to estimate the modeling systematic uncertainties. Both, simulated

events and data are processed using the same reconstruction algorithms and analysis chain.

7.4 Object Selection

The main physics objects considered in the analysis presented in this thesis are electrons, muons, jets and b -jets. The reconstruction, identification, isolation, and calibration definitions of these objects are described in Chapter 5. The different selection requirements of the physics objects are discussed below.

Events are required to have at least one vertex reconstructed from at least two tracks with p_T above 0.4 GeV. In case of several vertices, the one with the largest sum of the squared transverse momentum p_T of the associated tracks is taken.

Electrons are required to be central ($|\eta| < 2.47$), but outside the transition region between the barrel and the end-cap EM calorimeter ($1.37 < |\eta| < 1.52$), where a proper measurement is not possible, and to have $P_T > 10$ GeV. Electrons must pass a tight likelihood identification criterion (TightLH). Further selections on the longitudinal and transverse impact parameters, $|z_0 \sin\theta| < 0.5$ mm and $|\frac{d_0}{\sigma(d_0)}| < 5$, are imposed. These requirements are optimized to reduce the amount of fake and non-prompt electrons. Electrons must have p_T at least 1 GeV above the trigger threshold; above 25 (27) GeV depending on the 2015 (2016) triggers. This is done to avoid differences due to the calibration of the objects used in the trigger logic (online), and objects used in the physics analysis (off-line). To further reduce the background from non-prompt electrons coming from decays of hadrons in jets, electron candidates are also required to be isolated and to pass the *Gradient* isolation operating point which is tuned so that the electron-isolation efficiency is at least 90% for $p_T > 25$ GeV, increasing to 99% at 60 GeV, as detailed in Section 5.2.1.

Muons must satisfy *Medium* quality. This selection minimises the systematic uncertainties associated with muon reconstruction and calibration. Muons must have p_T above 25 (27) GeV depending on the 2015 (2016) triggers. Similar to electrons, they should also be isolated and pass the *Gradient* isolation requirement. The absolute value of a muon's d_0 significance must be less than 3, and the value of $|z \sin\theta|$ must be less than 0.5 mm.

Jets are reconstructed and calibrated to the particle level by the application of a jet energy scale (JES) derived from simulation and in situ corrections based on the 13 TeV

data. After energy calibration, jets are required to have $p_T > 25$ GeV and $|\eta| < 2.5$. Quality criteria, referred to as jet cleaning, are imposed to identify jets arising from non-collision sources or detector noise. As a result, events containing at least one such jet are removed. To reduce the effect of pileup and to avoid selecting jets from additional collisions within the same bunch crossing, an additional requirement is imposed on the tracks associated to the jet for low p_T ($p_T < 60$ GeV) jets in the central ($|\eta| < 2.4$) region of the detector. This algorithm is known as jet vertex tagger [130], referred to as JVT. A cut value of 0.59 [130] is assigned to identify jets which originate from the primary vertex.

b -jets are identified as originating from the hadronization of a b -quark (b -tagged) using multivariate techniques that combine information from the impact parameters of displaced tracks with the topological properties of secondary and tertiary decay vertices reconstructed within the jet. They are tagged via the MV2c10 [134] tagger, which is optimized to efficiently select jets containing b -hadron (b -jets) and separate them from jets containing c -hadrons (c -jets), jets containing hadronically decaying τ -leptons (τ -jets) and from other jets (light-jets). Four working points are defined by different MV2c10 discriminant threshold, as detailed in Table 5.2 in Chapter 5, referred to in the following as loose, medium, tight, very tight corresponding to a b -jet tagging efficiency of 85%, 77%, 70%, and 60%, respectively.

To resolve the potential ambiguities of a single detector response being assigned to two objects by the reconstruction algorithm, an overlap removal procedure is used. The energy deposits in the calorimeter are used to reconstruct electrons and jets. Therefore, double-counting of the electron energy deposits as a jet could occur. To prevent this, the closest jet within $\Delta R_y = \sqrt{\Delta y^2 + \Delta \phi^2} = 0.2$ of a selected electron is removed³. If the nearest jet surviving that selection is found within $\Delta R_y = 0.4$ of the electron, the electron is discarded. In the case of muons, they are removed if they are separated from the nearest jet by $\Delta R_y < 0.4$. This reduces the background arising from heavy-flavor decays inside jets. However, if the jet has less than three associated tracks, the muon is kept and the jet is removed instead. This is done to prevent an inefficiency for high-energy muons undergoing significant energy loss in the calorimeter. A hadronic τ candidate is rejected if it is separated by $\Delta R_y < 0.2$ from any selected electron or muon.

3. The rapidity is defined as $y = \frac{1}{2} \times \ln \frac{E+p_z}{E-p_z}$, where E is the energy and p_z is the longitudinal component of the momentum along the beam pipe.

7.5 Event Selection and Categorization

First, events containing top-quark pair production are selected as described in Section 7.5.1, then they are categorized in different regions enhanced in the $t\bar{t}H$ signal including background processes. The categorization is done in order to improve the sensitivity of the search and to create regions that are pure with specific processes such as the $t\bar{t}H$ signal and the dominant backgrounds. The first part of the categorization uses the particle content of the additional jets of the event, as described in Section 7.5.2. The second part uses the information of how likely reconstructed events are to contain b -jets, as detailed in Section 7.5.3.

7.5.1 Event Selection of Recorded Data

Events with single lepton top-quark decays are selected, containing exactly one lepton with a p_T above 27 GeV. In the single-lepton channel events are required to contain at least five jets, of which two must be b -tagged using the loosest operating point.

An additional search category is being considered in this analysis which will be briefly discussed here. This category targets events in which the Higgs boson and top-quarks are produced with high transverse momenta (boosted), such that their decay products are more collimated and can be reconstructed within a single large radius jet; with $R = 1.0$. Hence, referred to as the boosted category. More details on the boosted category can be found in Appendix A.1.

Some of the selected $t\bar{t}H$ candidates in the single-lepton channel also pass the boosted selection requirements. However, in favor of the boosted category and to avoid overlapping, these events are removed from the resolved single-lepton channel and added to the boosted category. About 1% of the $t\bar{t}H$ and $< 0.1\%$ $t\bar{t}$ +jets events are removed from the inclusive sample.

Events in the dilepton channel must have at least three jets in which two of them must be b -tagged using the medium operating point, and exactly two leptons with opposite-sign electric charge. In all the considered selections, at least one reconstructed lepton with p_T above 27 GeV is required to match within $\Delta R < 0.15$ to a lepton with the same flavor reconstructed by the trigger algorithm. The p_T of the sub-leading lepton must be above 15 GeV in the ee channel or above 10 GeV in the $e\mu$ or $\mu\mu$ channels. The dilepton invariant

mass in the ee and $\mu\mu$ channels must be above 15 GeV and outside the Z mass window 83 – 99 GeV.

In order to avoid overlapping with searches in other $t\bar{t}H$ decay modes such as the multi-lepton channel [184], events are removed if they contain a hadronic τ with a p_T above 25 GeV. Events which fail the dilepton channel requirements and contain exactly one lepton with p_T above 27 GeV enter the single-lepton channel. However events in the single-lepton channel which contain at least two hadronic τ leptons with p_T above 25 GeV are removed.

After the analysis selection described above, the data sample is dominated by background from $t\bar{t}$ events. In addition to the main background, there are small contributions from the associated production of a vector boson and a $t\bar{t}$ pair ($t\bar{t} + V$; $V = W, Z$) and non- $t\bar{t}$ events such as the production of a single top, followed by the production of a W or a Z boson in association with jets (W/Z +jets), diboson (WW, WZ, ZZ), fake and non-prompt leptons. Backgrounds from non- $t\bar{t}$ processes are grouped together in the figures and represented in yellow, unless stated otherwise.

7.5.2 Event Categorization at Particle Level

The $t\bar{t}$ +jets events are classified into three non-overlapping samples according to the flavor of the additional jets that do not originate from the decay of the $t\bar{t}$ system. Particle jets are reconstructed from stable⁴ truth particles, as described in Section 5.3, using the anti- k_t jet finding algorithm with a radius parameter $R = 0.4$. Particle jets are required to have $p_T > 15$ GeV and $|\eta| < 2.5$. Events are labelled as $t\bar{t}+ \geq 1b$ if at least one particle jet is matched within $\Delta R < 0.4$ to a b -hadron with $p_T > 5$ GeV not originating from the decay of a top-quark. Similarly, if at least one particle jet is matched to a c -hadron, which is not a decay product of a b -hadron, with $p_T > 5$ GeV not originating from W boson, the event is labelled as $t\bar{t}+ \geq 1c$. Events that are labelled as either $t\bar{t}+ \geq 1b$ or $t\bar{t}+ \geq 1c$ are referred to as $t\bar{t} + HF$, where HF stands for heavy-flavor. The remaining events including those with no additional jets are labelled as $t\bar{t}$ +light-jets.

4. Stable refers to a final-state particle with mean lifetime $\tau = 3 \times 10^{-11}$ s.

7.5.3 Event Categorization at Reconstruction Level

The categorization of events is designed to define regions of phase space, obtained by applying a selection on the number of jets and b -tagged jets, where the signal model predicts a significant excess of events over the predicted background level, S/B and S/\sqrt{B} ("S" refers to the number of events of the SM Higgs boson signal, and "B" refers to the expected number of background events according to MC simulation). Such regions are referred to as signal enriched regions or signal regions (SR). To estimate background processes contaminating the signal regions, one typically defines control regions (CR), in which the dominant backgrounds can be controlled by comparison to the data samples. Control regions are specifically designed to have a high purity of one type of background. Using the four working points of the MV2c10 tagger, regions rich in $t\bar{t}H$ signal and the main backgrounds, $t\bar{t} + b$, $t\bar{t} + c$ and $t\bar{t}$ +light, are determined.

Events in the resolved single-lepton (dilepton) channel are first classified depending on whether the number of jets is five (three) or at least six (four). Figure 7.5 shows the definition of the five-jet and six-jet signal and control regions for the resolved single-lepton channel depending on the b -tagging requirements. The y -axis defines the b -tagging operating point of the first and second b -tagged jets. While, the x -axis defines those for the third and fourth b -tagged jets. Regions with similar background composition are merged together, resulting in 11 regions. The dilepton channel has 7 regions, and is detailed in Appendix A.2.

The signal regions are defined with different levels of purity of the $t\bar{t}H$ signal and $t\bar{t} + b\bar{b}$ background components. The purest signal regions are $SR_1^{\geq 6j}$ and SR_1^{5j} , which require four b -tagged jets with the very-tight operating point at 60%. Looser requirements are imposed to the other signal regions referred to as $SR_2^{\geq 6j}$, $SR_3^{\geq 6j}$, and SR_2^{5j} . Events passing the boosted single-lepton channel selection form the sixth signal region SR^{boosted} . The remaining events with exactly five jets are categorized in three control regions enriched in $t\bar{t} + b$, $t\bar{t} + \geq 1c$, and $t\bar{t}$ +light, referred to as $CR_{t\bar{t}+b}^{5j}$, $CR_{t\bar{t}+\geq 1c}^{5j}$, and $CR_{t\bar{t}+\text{light}}^{5j}$, respectively. Similarly, remaining events with at least six jets form the other three control regions, referred to as $CR_{t\bar{t}+b}^{\geq 6j}$, $CR_{t\bar{t}+\geq 1c}^{\geq 6j}$, and $CR_{t\bar{t}+\text{light}}^{\geq 6j}$.

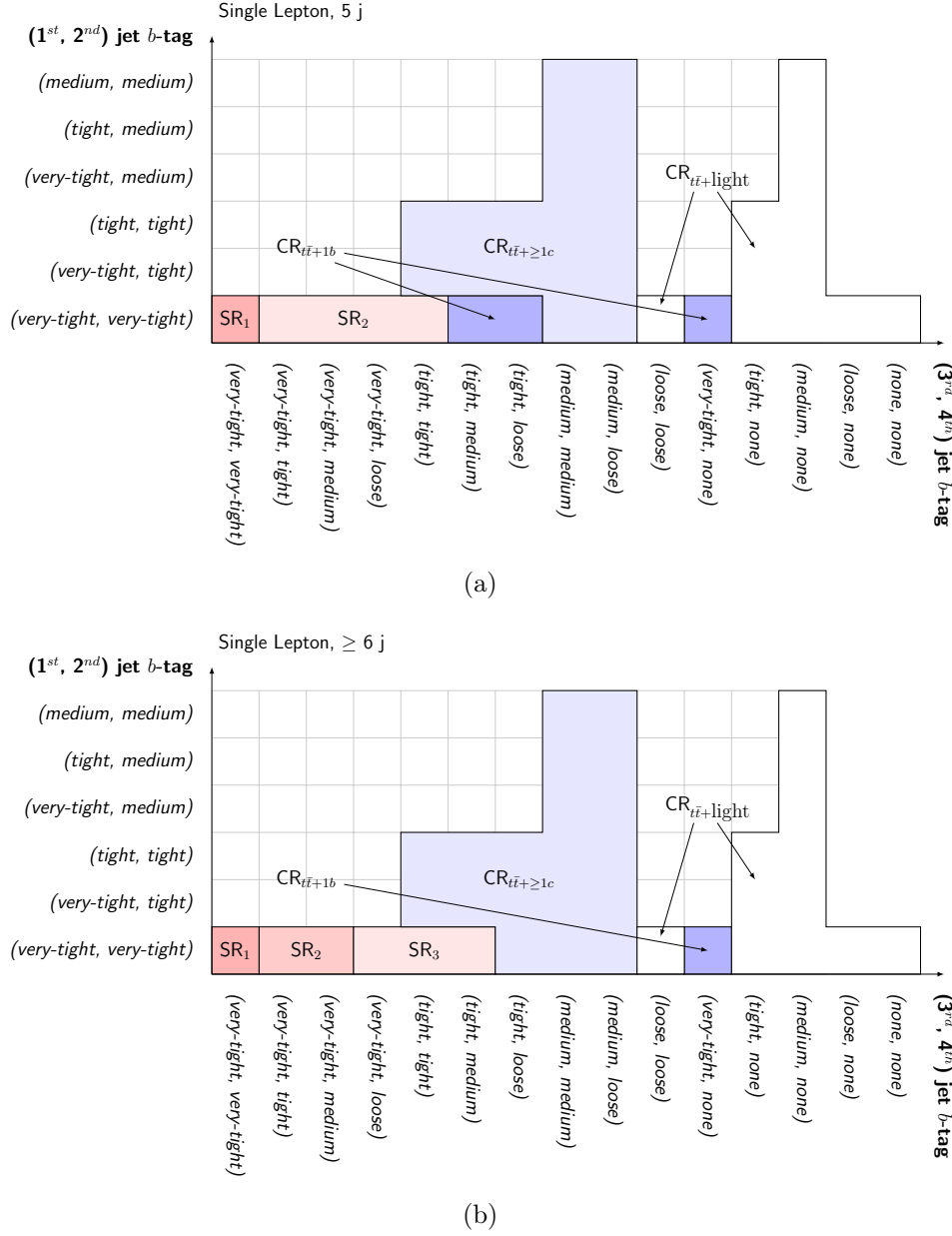


Figure 7.5: Definition of the (a) five-jet and (b) six-jet signal and control regions in the resolved single-lepton channel, based on jets ordered in terms of b -tagging operating points. The vertical axis shows the requirements on the first two jets, while the horizontal axis on the third and fourth jets. The jets are ordered such that the ones passing tighter b -tagging requirements are considered first, which means the empty squares are not possible.

The signal purity for each of the signal and control regions in the single-lepton and dilepton channel is shown in Figure 7.6 and Figure 7.7, respectively. Despite the effort to enhance the $t\bar{t}H$ events in the signal region, the most sensitive regions in the single-lepton (dilepton) channels have $S/B = 5.3\%$ ($S/B = 5.4\%$).

The expected proportions of different backgrounds in each region are shown in Figure 7.8 and Figure 7.9. The main background contribution in the signal regions arises from $t\bar{t} + \geq 1b$, about 80% in $SR_1^{\geq 6j}$ and $SR_1^{\geq 4j}$. In total four control regions are dominated by $t\bar{t}$ +light background, three control regions dominated by $t\bar{t} + \geq 1c$ background, two control regions dominated by $t\bar{t} + b$ background, and one control region dominated by $t\bar{t} + \geq 1b$ background. The $t\bar{t} + \geq 1c$ control regions have a large contamination of $t\bar{t}$ +light and $t\bar{t} + \geq 1b$ background since the MV2c10 tagger is developed to discriminate b -jets from c - or light-jets but not to discriminate c -jets from light-jets.

$\sqrt{s} = 13 \text{ TeV}, 36.1 \text{ fb}^{-1}$

Single Lepton

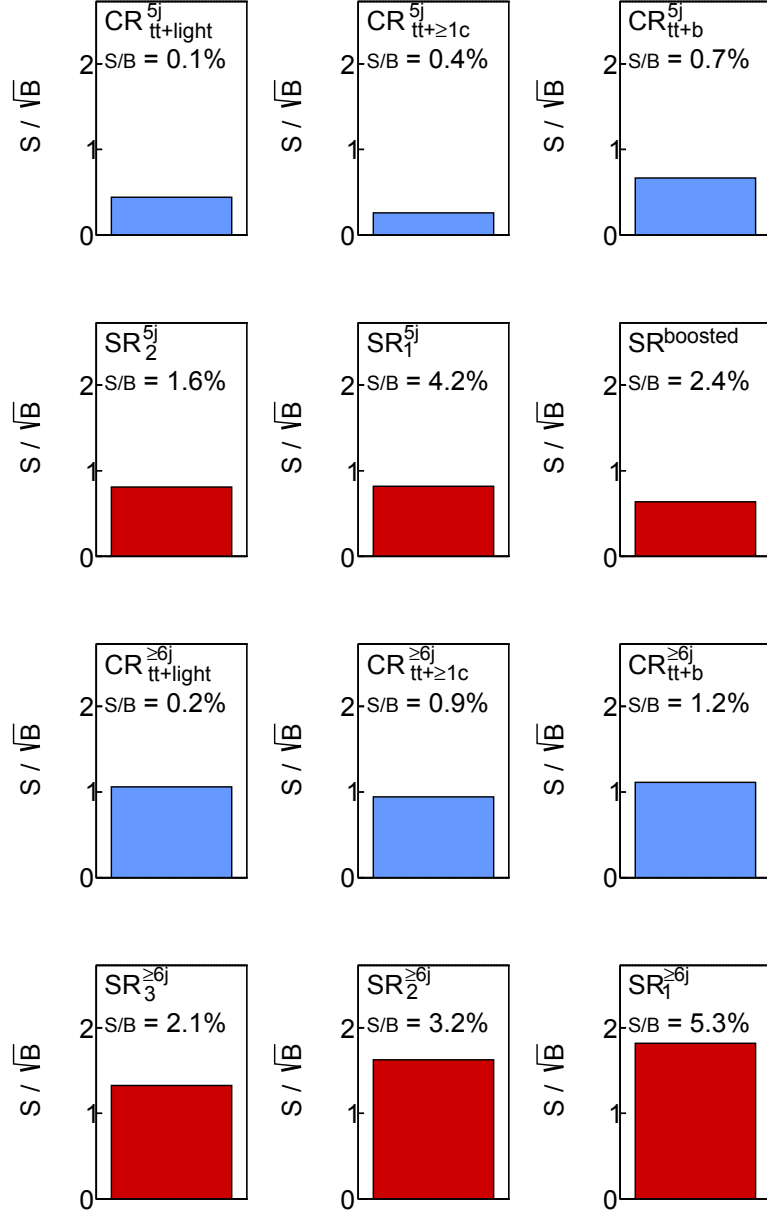


Figure 7.6: Analysis regions for the single-lepton channel. The S/\sqrt{B} and S/B ratios for each of the regions are shown where S (B) is the number of selected signal (background) events. Signal regions are shaded in red, while the control regions are shown in blue.

$\sqrt{s} = 13 \text{ TeV}, 36.1 \text{ fb}^{-1}$
Dilepton

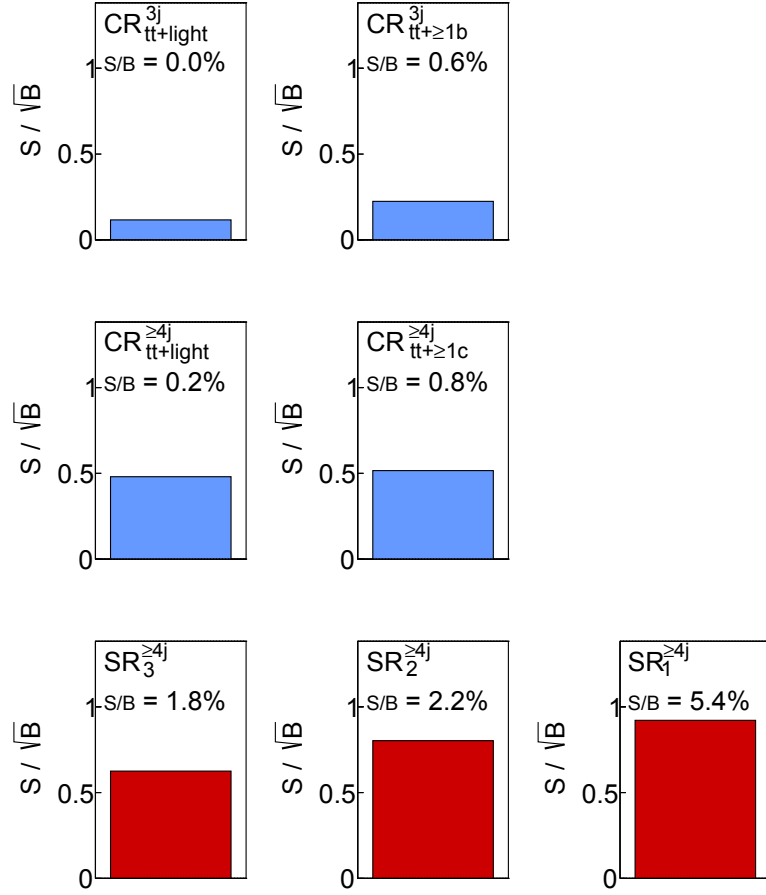


Figure 7.7: Analysis regions for the dilepton channel. The S/\sqrt{B} and S/B ratios for each of the regions are shown where S (B) is the number of selected signal (background) events. Signal regions are shaded in red, while the control regions are shown in blue.

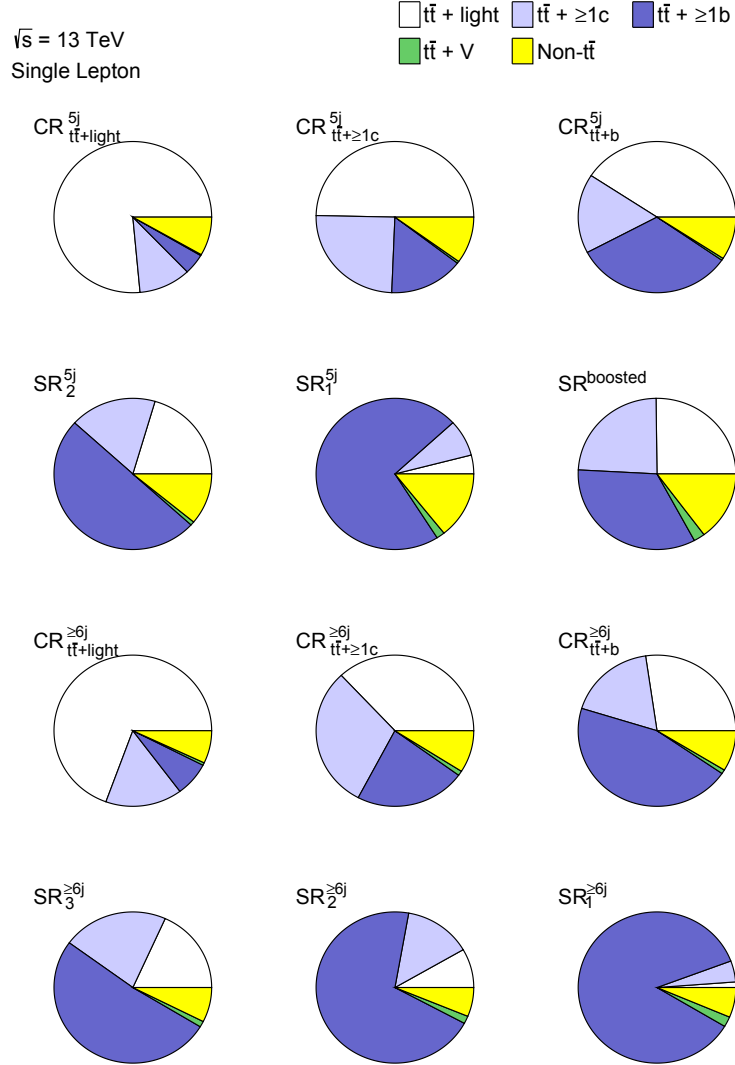


Figure 7.8: Fractional contributions of the various backgrounds to the total background prediction in each analysis category in the single-lepton channel. See text for details about the categorization of the $t\bar{t}$ background. Backgrounds from non- $t\bar{t}$ processes are grouped together and represented in yellow.

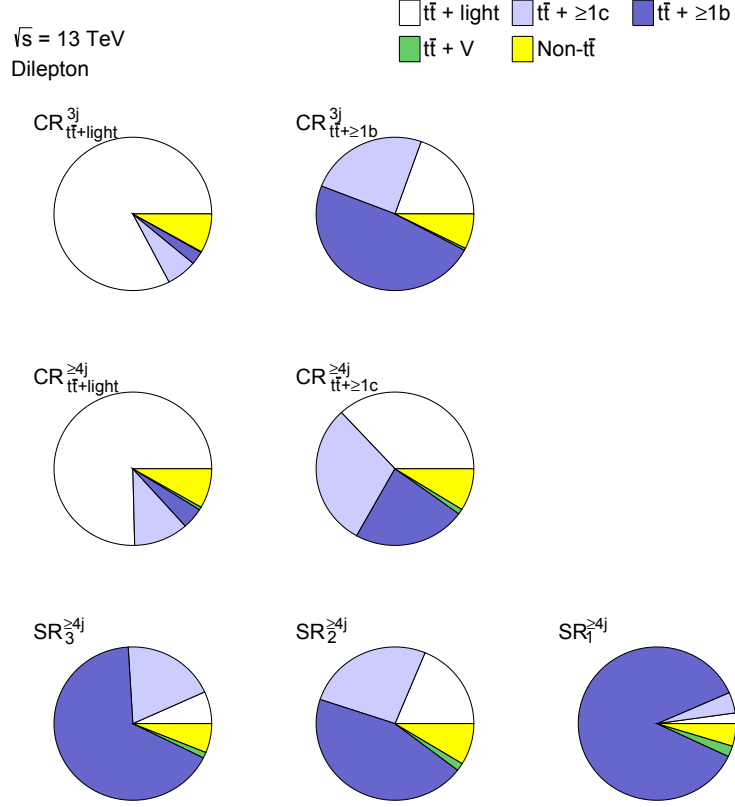


Figure 7.9: Fractional contributions of the various backgrounds to the total background prediction in each analysis category in the dilepton channel. See text for details about the categorization of the $t\bar{t}$ background. Backgrounds from non- $t\bar{t}$ processes are grouped together and represented in yellow.

7.6 Multivariate Analysis

The small signal-to-background ratio in the signal regions after event selection and categorization, and the fact that the $t\bar{t}H(H \rightarrow b\bar{b})$ signal has similar final states compared to the dominating background arising from $t\bar{t} + \geq 1b$ events require additional steps to separate the signal from background. Reconstruction of the Higgs boson mass peak is not possible due to the large combinatorial background arising from the presence of four b -jets in the final state. Also, the absence of a single variable that can exhibit a clear separation power among signal and background events enhances the necessity to use multivariate analysis (MVA) approach in order to better distinguish signal events from the background. The use of MVA aims to maximize the amount of information and explores correlations among different variables to distinguish signal events from background ones.

Two different steps are used. The first one is dedicated to the reconstruction of the Higgs boson from the final-state partons of the $t\bar{t}H(H \rightarrow b\bar{b})$ system, as described in Section 7.6.2. The second step is used to separate the $t\bar{t}H$ signal from the main background originating from $t\bar{t}$ +jets, as detailed in Section 7.6.3.

7.6.1 Boosted Decision Trees

Boosted Decision Trees (BDT) are a set of binary structured decision trees that use the *boosting* technique. BDTs are among the most used machine learning techniques in high energy physics.

Decision trees (DT) were formalized and developed by Breiman [185] in the context of pattern recognition and data mining. They extend a simple cut-based analysis into a multivariate technique by continuing to analyse events that fail a particular criterion until they satisfy a terminating condition [186] as follows.

BDTs are trained using decision trees, which are binary tree networks for data categorization that classify events between signal and background, as shown in Figure 7.10. The decision tree starts from a root node that contains all the events, and grows successive layers formed by binary nodes. At each node, a cut on a particular discriminating variable " x_i " is applied to split the data. When a new node is generated, the variable and the cut value that can achieve the best separation among signal and background, is selected. Each event starts from the root node and goes down the decision tree. The output of the DT,

referred to as the end-node's signal purity or leaves, classifies the events more signal-like with values close to 1 or more background-like with values close to 0.

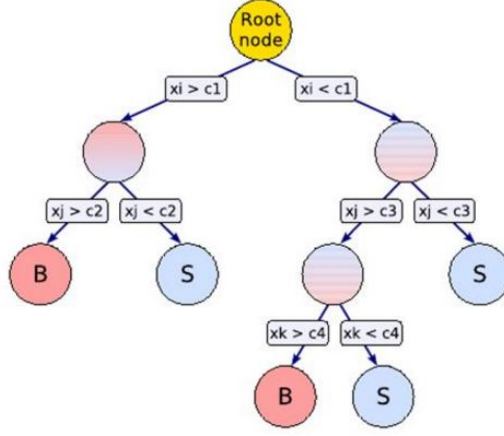


Figure 7.10: Schematic view of a decision tree. Starting from the root node, a sequence of binary splits using the discriminating variables (x) are split into smaller regions [187]. The cut values and the order of the nodes are determined using a so called training algorithm.

The intention during the training is to achieve the most optimal split, S^* , among signal and background. This is done after scanning all variables for all the events at each node. Then, S^* is selected from all splits (S), as the one maximizing the decrease of impurity, $\Delta i(S^*)$ according to

$$\Delta i(S^*) = \max_{S \in \text{splits}} \Delta i(S), \quad (7.2)$$

where $\Delta i(S)$ is defined as

$$\Delta i(S) = i - \min[p_P i_P, p_F i_F], \quad (7.3)$$

where p_P (p_F) is the fraction of events passing (failing) the split (S), i stands for the impurity, and i_P (i_F) is the impurity for passing (failing) events. The impurity definition in DTs is defined as:

$$i = 2p \cdot (1 - p), \quad (7.4)$$

which is referred to as *Gini-index*, in which p is the signal purity ($\frac{s}{s+b}$) and s (b) is

the number of signal (background) events.

A single decision tree is limited in its separating power. Therefore, various decision trees are combined together to form a "forest". A sequence of trees are generated by the algorithm where more emphasis is given to the previously misclassified events. This process of assembling many DT and combining them together to form a single strong classifier is called "boosting". In the analysis presented in this thesis, the Adaptive Boost (AdaBoost) [188] algorithm is used. This algorithm starts with the original event weights when training the first decision tree. Then, the subsequent tree is trained using a modified event sample, in which the weights of the previously misclassified events are multiplied by a common "boost" weight. After processing all decision trees, the event is classified as signal or background depending on the weighted average of the individual tree classifications. This weighted average is the final BDT score which is the likelihood of an event to be a signal or background, as shown in Figure 7.11 (a). The studies presented in this thesis are based on boosted decision trees (BDT) which are trained using the Toolkit for Multivariate Analysis (TMVA) [187].

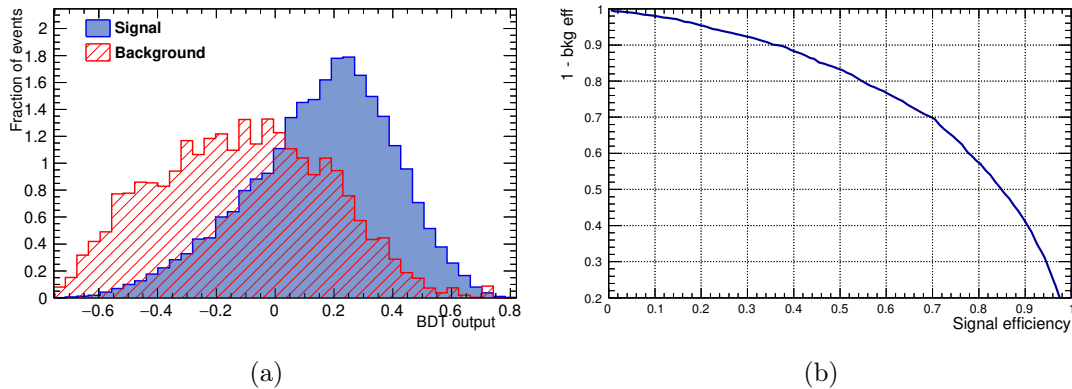


Figure 7.11: (a) BDT output score showing the signal events in blue versus the background events in red. The BDT output classifies the events more signal-like with values close to 1 or more background-like with values close to 0. Corresponding to events in (a), Receiver Operating Characteristic (ROC) curve (b) is generated by the TMVA framework. Background rejection ($1 - \text{bkg eff}$) is shown as a function of the signal efficiency. The ratio of selected signal events to the total signal events is known as the signal efficiency, and the ratio of rejected background events to the total background events is known as the background efficiency.

In order to evaluate the performance of each trained BDT, receiver operating characteristic (ROC) curves are used. These curves illustrate background rejection versus signal efficiency caused by a variation of the threshold on the BDT score, as shown in Figure 7.11 (b). The signal efficiency is defined as the proportion of signal events above a particular threshold on the BDT score to all signal levels. While, background rejection is defined as $1 - \text{background efficiency}$ (referred to as $(1 - \text{bkg eff})$ in Figure 7.11 (b)), which is the proportion of rejected background by the same threshold. Better BDT performance means higher background rejection at similar signal efficiency, resulting in a more convex ROC curve.

In the $t\bar{t}H$ analysis presented here, the full spectrum of the BDT output score is used. Therefore, the total area under the ROC curve (AUROC) represents the separating performance of a particular trained BDT.

Overtraining is a major concern for the BDT training. If it occurs, the BDT describes statistical fluctuations in the data set used for the training leading to a performance that does not hold if data sets are changed. In order to avoid this, the training sample of MC simulated data is split in two set of statistically independent samples (sample A and sample B) based on the event number. The BDT trained on even events is then applied on odd events, and vice versa, referred to as *cross training*. Cross training profits from the full available statistics by evaluating the events in sample B with the BDT trained on sample A and the opposite.

The BDTs used here are constructed from 400 individual trees, with a maximum depth of 5 nodes. The BDT is first trained on MC simulated data that contain both signal and background, and then applied to data assuming that the same separation power holds. This is justified if all the used variables in the training are well modeled in MC compared to data.

7.6.2 MVA-based Reconstruction of the $t\bar{t}H$ Final State

A full event reconstruction using a BDT is performed in all single-lepton and dilepton signal regions. For simplicity, the method will be described in detail for the single-lepton channel; a similar procedure is used in the dilepton channel. The Reconstruction BDT is trained on $t\bar{t}H$ signal events to correctly assign the reconstructed jets to the final state

partons from top-quark and Higgs boson decays, and to suppress background from wrong combinations. For this reason, W -boson, top-quark and Higgs-boson candidates are built from reconstructed jets, missing transverse energy and one lepton. Jets are assigned to the quarks from the $t\bar{t}H(H \rightarrow b\bar{b})$ decay and combinations including jets and b -jets are used to reconstruct the objects such as the Higgs boson and the hadronic top-quark. The following summarizes the used reconstruction algorithm.

- **Reconstruction of the leptonic W boson**

The leptonic W boson is reconstructed using the lepton and the neutrino four-momenta. The transverse momentum of the neutrino can be measured using the missing transverse momentum (E_T^{miss}), but the longitudinal component of the neutrino momentum, $p_{z\nu}$, is not measurable. It can be inferred by assuming the lepton and E_T^{miss} are originating from the W boson decay. Therefore, the sum of the lepton and neutrino four momenta is equal to the four momentum of the W boson. Using the expected W boson mass ($M_W = 80.385 \text{ GeV}$) [20], one can compute $p_{z\nu}$. This leads to a quadratic equation with two possible solutions:

$$p_{z\nu}^{\pm} = \frac{1}{2} \frac{p_{z\ell}\beta \pm \sqrt{\Delta}}{E_{\ell}^2 - p_{z\ell}^2} \quad (7.5)$$

where:

$$\beta = m_W^2 - m_{\ell}^2 + 2p_{x\ell}p_{x\nu} + 2p_{y\ell}p_{y\nu} \quad (7.6)$$

$$\Delta = E_{\ell}^2(\beta^2 + (2p_{z\ell}p_{T\nu})^2 - (2E_{\ell}p_{T\nu})^2). \quad (7.7)$$

If no real solutions exist, the discriminant of the quadratic equation is set to zero ($\Delta = 0$), giving a unique solution. In the case of two solutions, two different leptonic W bosons are considered in a separate contribution.

- **Reconstruction of the hadronic W boson**

The hadronic W boson candidates are reconstructed using all combinations of two jets which are not b -tagged. In the five-jet categories the sub-leading quark from the W boson is not matched in most of the events. Therefore, the hadronic W boson is not reconstructed for events with exactly five jets.

- **Reconstruction of the top-quarks and the Higgs boson**

The Higgs boson candidates are reconstructed from all possible pairs of b -jets. While, the top-quark candidates are reconstructed from one W -boson candidate and one b -tagged jet. The top-quark candidate containing the leptonically (hadronically) decaying W boson is referred to as the leptonically (hadronically) decaying top-quark candidate.

The $t\bar{t}H$ signal sample is used for training the BDT. All combinations of possible jet/lepton to parton assignment are considered, and all described kinematic variables are computed. The correctly assigned combination is considered as "signal" and the incorrect combinations are considered as "background" in the BDT training. Under the training, the BDT thus establishes a correlation between correct jet-parton matching and its related kinematic properties. The trained BDT is then applied to data events, where all combinations of jet assignments to either the Higgs boson or top-quark decay are constructed and ordered by the BDT score. The correct combination is the one with the highest BDT score, which is then used to calculate the invariant mass and angular separations in addition to other kinematic variables, which feed into the Classification BDT training.

A maximum of 19 kinematic variables depending on the region, are built as input variables for the BDT. The variables are listed in Table 7.4. They are chosen to address particular kinematic characteristics of the correct and wrong jet combinations.

The best reconstruction performance can be obtained by including information related to the Higgs boson such as the predicted Higgs boson invariant mass. However, when this BDT is run on data which contain both $t\bar{t}H$ signal and $t\bar{t}$ +jets background events, the use of the Higgs mass will bias the selection of a particular b -jet combination to be more Higgs-like and reduces the ability to separate signal from background. Therefore, two

Variable	$SR_{1,2,3}^{\geq 6j}$	$SR_{1,2}^{5j}$
Topological information from $t\bar{t}$		
Mass of top_{lep}	✓	✓
Mass of top_{had}	✓	–
Mass of q_1 from W_{had} and b from top_{had}	–	✓
Mass of W_{had}	✓	–
Mass of W_{had} and b from top_{lep}	✓	–
Mass of q_1 from W_{had} and b from top_{lep}	–	✓
Mass of W_{lep} and b from top_{had}	✓	✓
$\Delta R(W_{\text{had}}, b \text{ from } \text{top}_{\text{had}})$	✓	–
$\Delta R(q_1 \text{ from } W_{\text{had}}, b \text{ from } \text{top}_{\text{had}})$	–	✓
$\Delta R(W_{\text{had}}, b \text{ from } \text{top}_{\text{lep}})$	✓	–
$\Delta R(q_1 \text{ from } W_{\text{had}}, b \text{ from } \text{top}_{\text{lep}})$	–	✓
$\Delta R(\ell, b \text{ from } \text{top}_{\text{lep}})$	✓	✓
$\Delta R(\ell, b \text{ from } \text{top}_{\text{had}})$	✓	✓
$\Delta R(b \text{ from } \text{top}_{\text{lep}}, b \text{ from } \text{top}_{\text{had}})$	✓	✓
$\Delta R(q_1 \text{ from } W_{\text{had}}, q_2 \text{ from } W_{\text{had}})$	✓	–
$\Delta R(b \text{ from } t_{\text{had}}, q_1 \text{ from } W_{\text{had}})$	✓	–
$\Delta R(b \text{ from } t_{\text{had}}, q_2 \text{ from } W_{\text{had}})$	✓	–
Min. $\Delta R(b \text{ from } \text{top}_{\text{had}}, q_i \text{ from } W_{\text{had}})$	✓	–
$\Delta R(\text{lep}, b \text{ from } \text{top}_{\text{lep}}) - \min. \Delta R(b \text{ from } \text{top}_{\text{had}}, q_i \text{ from } W_{\text{had}})$	✓	✓
Topological information from the Higgs-boson candidate		
Mass of Higgs	✓	✓
Mass of Higgs and q_1 from W_{had}	✓	✓
$\Delta R(b_1 \text{ from Higgs}, b_2 \text{ from Higgs})$	✓	✓
$\Delta R(b_1 \text{ from Higgs}, \text{lepton})$	✓	✓
$\Delta R(b_1 \text{ from Higgs}, b \text{ from } \text{top}_{\text{lep}})$	–	✓
$\Delta R(b_1 \text{ from Higgs}, b \text{ from } \text{top}_{\text{had}})$	–	✓

Table 7.4: Input variables to the Reconstruction BDT in the single lepton channel. The subscript had(lep) indicates the hadronically (leptonically) decaying W or t and q_i refers to quarks from W . $SR_{1,2,3}^{\geq 6j}$, and $SR_{1,2}^{5j}$ correspond to the six- and five-jet resolved single-lepton regions defined in Figure 7.5.

versions of the Reconstruction BDT are used in each signal-enriched region; one with and one without the Higgs boson information, and both jet-parton assignments are considered when computing input variables for the Classification BDT.

The performance of the MVA-based reconstruction is quantified using the reconstruction efficiency which is defined as the fraction of events for which the chosen combination is the correct one. For example, the Higgs boson is correctly reconstructed in 48% (30%) of the selected $t\bar{t}H$ events in the most sensitive signal region $SR_1^{\geq 6j}$ using the Reconstruction BDT with (without) kinematic information about the Higgs boson.

Two other intermediate MVA techniques, which are briefly mentioned here and are detailed in [150], are used. The first method, is a Likelihood Discriminant (LHD), which

provides a single likelihood discriminant for each event to satisfy the $t\bar{t}H(H \rightarrow b\bar{b})$ signal or $t\bar{t}$ +jets background hypotheses. The probability for each event to be signal or background is computed using reference distributions for the kinematics of the final state objects. Then, the probabilities are combined within a likelihood discriminant. The second one, is the Matrix Element Method (MEM) that is based on the integration of the matrix element for each event assuming the signal or background Feynman diagrams at leading order.

7.6.3 Discrimination between Signal and Background

The Classification BDT is used to classify events as more signal or background-like. It is built upon input variables that exploit various kinematic differences of the signal and background events, as well as the b -tagging information. The general kinematic variables are the invariant masses and the angular separations of pairs of reconstructed jets and leptons, and the b -tagging discriminant of the selected jets. This is combined with both Reconstruction BDT and the outputs of the LHD and MEM. Four sets of variables, which are listed in Table 7.5, are used as inputs to a Classification BDT that provides the final discrimination between the $t\bar{t}H$ signal and the $t\bar{t}$ +jets background.

About 30 variables are selected for the starting point on the basis of their discrimination power. Then, an interactive process is used to find an optimal set of variables in each signal region. The input variables are ranked by their signal-to-background separation power through the TMVA separation, which is defined as:

$$\frac{1}{2} \sum_i^{\text{bins}} \frac{(N_i^S - N_i^B)^2}{N_i^S + N_i^B}, \quad (7.8)$$

where N_i^S and N_i^B are the entries in each bin of the normalized signal and background histograms, respectively. Variables that show no significant improvement of discrimination between signal and background are removed. In the end, only 19 variables are selected in the six-jet regions and 18 variables are selected in the five-jet regions, as shown in Table 7.5.

Variable	Definition	$SR_{1,2,3}^{\geq 6j}$	$SR_{1,2}^{5j}$
General kinematic variables			
$\Delta R_{bb}^{\text{avg}}$	Average ΔR for all b -tagged jet pairs	✓	✓
$\Delta R_{bb}^{\text{max } p_T}$	ΔR between the two b -tagged jets with the largest vector sum p_T	✓	–
$\Delta \eta_{jj}^{\text{max } \Delta \eta}$	Maximum $\Delta \eta$ between any two jets	✓	✓
$m_{bb}^{\text{min } \Delta R}$	Mass of the combination of two b -tagged jets with the smallest ΔR	✓	–
$m_{jj}^{\text{min } \Delta R}$	Mass of the combination of any two jets with the smallest ΔR	–	✓
N_{30}^{Higgs}	Number of b -jet pairs with invariant mass within 30 GeV of the Higgs boson mass	✓	✓
H_T^{had}	Scalar sum of jet p_T	–	✓
$\Delta R_{\ell,bb}^{\text{min } \Delta R}$	ΔR between the lepton and the combination of the two b -tagged jets with the smallest ΔR	–	✓
Aplanarity	$1.5\lambda_2$, where λ_2 is the second eigenvalue of the momentum tensor [99] built with all jets	✓	✓
$H1$	Second Fox–Wolfram moment computed using all jets and the lepton	✓	✓
Variables from reconstruction BDT			
BDT output	Output of the reconstruction BDT	✓*	✓*
m_{bb}^{Higgs}	Higgs candidate mass	✓	✓
$m_{H,b\text{lep top}}$	Mass of Higgs candidate and b -jet from leptonic top candidate	✓	–
$\Delta R_{bb}^{\text{Higgs}}$	ΔR between b -jets from the Higgs candidate	✓	✓
$\Delta R_{H,t\bar{t}}$	ΔR between Higgs candidate and $t\bar{t}$ candidate system	✓*	✓*
$\Delta R_{H,\text{lep top}}$	ΔR between Higgs candidate and leptonic top candidate	✓	–
$\Delta R_{H,b\text{had top}}$	ΔR between Higgs candidate and b -jet from hadronic top candidate	–	✓*
Variables from Likelihood and Matrix Element Method calculations			
LHD	Likelihood discriminant	✓	✓
MEM_{D1}	Matrix Element discriminant	✓	–
Variables from b -tagging			
$w_{b\text{-tag}}^{\text{Higgs}}$	Sum of b -tagging discriminants of jets from best Higgs candidate from the reconstruction BDT	✓	✓
B_{jet}^3	3 rd largest jet b -tagging discriminant	✓	✓
B_{jet}^4	4 th largest jet b -tagging discriminant	✓	✓
B_{jet}^5	5 th largest jet b -tagging discriminant	✓	✓

Table 7.5: Input variables to the classification BDT in the single-lepton channel. For variables from the reconstruction BDT, those with a * are from the BDT using Higgs boson information, while those with no * are from the BDT without Higgs boson information. The MEM_{D1} variable is only used in the $SR_1^{\geq 6j}$.

Each individual input variable to the Classification BDT shows only small kinematic differences between signal and background, as shown in Figure 7.12. Figure 7.12 (a) shows the average opening angle between all b -tagged jet pairs ($\Delta R_{bb}^{\text{avg}}$). The b -tagged jet pairs originating from the Higgs boson (in blue in Figure 7.12 (a)) are more collimated than the b -tagged jet pairs originating from $t\bar{t}$ +jets background (in red). Figure 7.12 (b) shows the invariant mass of the Higgs boson candidate where the signal events have a clear peak around the Higgs boson mass compared to the background events. Figure 7.12 (c) shows that more b -jet pairs with invariant mass within 30 GeV of the Higgs boson mass are expected for the signal events compared to the background ones.

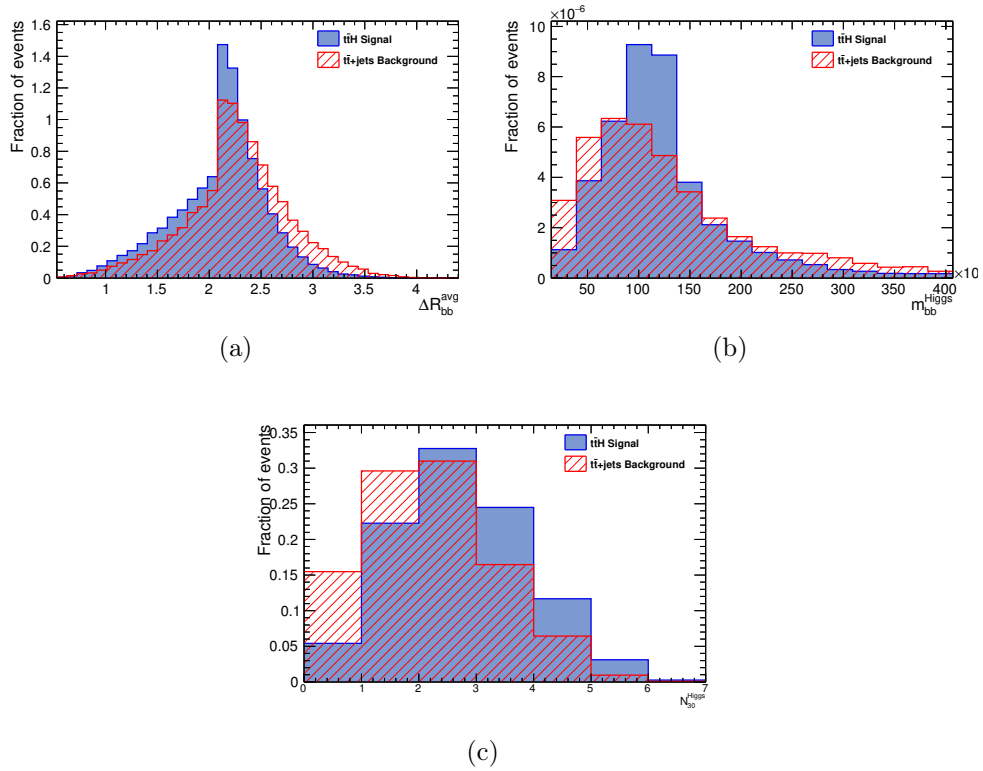
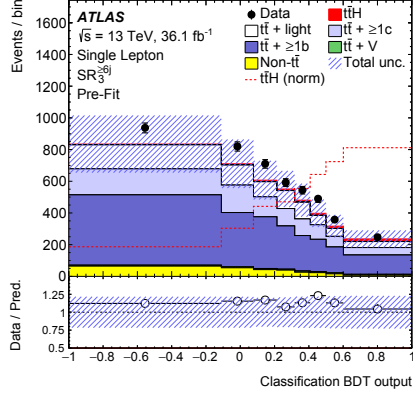


Figure 7.12: Examples of training variables for the Classification BDT in $SR_1^{\geq 6j}$, showing distributions for signal and background of (a) the average opening angle between all b -tagged jet pairs, (b) the invariant mass of the Higgs candidate, and (c) the number of b -jet pairs with invariant mass within 30 GeV of the Higgs boson mass.

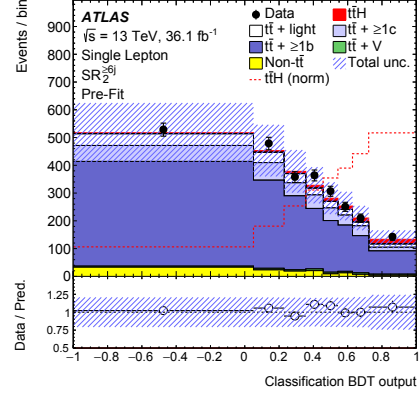
The above mentioned variables are combined in the BDT discriminant that shows the best separation power compared to the input variables, as shown in Figure 7.13. Therefore, it is used as the final discriminants in the fit to data in the signal regions. Figure 7.13 shows

the comparison of data and MC prediction for the distributions of the Classification BDT discriminant in the six-jet signal regions. The dashed red line in Figure 7.13 shows the $t\bar{t}H$ signal distribution normalised to the total background prediction, for better visibility of the shape of the Higgs boson signal. The low bins of the classification BDT output has almost no signal events, whereas the last three bins show high signal events. The BDT helps to increase the overall signal to background (S/B) ratio from 5.3% in $SR_1^{\geq 6j}$ to about 20% in the last bin of the BDT score.

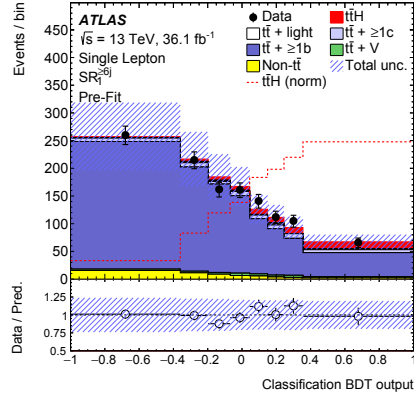
Higher yields are observed in data compared to MC simulation, with variations up to 15% as shown in Figure 7.13 (a). The normalization of the $t\bar{t}+ \geq 1b$ and $t\bar{t}+ \geq 1c$ backgrounds are not included in the assigned uncertainties and are determined from the fit to data, which will be discussed in Section 7.10.2. However, the differences seen between data and MC are within the assigned uncertainties. The data is well described by the simulated MC which is a mixture of the various background processes.



(a)



(b)



(c)

Figure 7.13: Comparison between data and prediction for the classification BDT output distributions in the six-jet signal regions in the single lepton channel. The dashed red line in shows the $t\bar{t}H$ signal distribution normalised to the total background prediction. The uncertainty band contains both statistical uncertainty and systematic uncertainties. Distributions are shown before the fit procedure, uncertainties on the normalisation of $t\bar{t} + \geq 1b$ or $t\bar{t} + \geq 1c$ are not included.

7.7 Background Estimation

The challenge of the $t\bar{t}H(H \rightarrow b\bar{b})$ analysis is the large amount of background arising from $t\bar{t}$ +jets processes that have a significantly higher cross section than the signal. Therefore, precise background estimation is crucial and described in Section 7.7.1. Primary focus is placed on the $t\bar{t}+ \geq 1b$ background which is the dominant background in the signal regions, as detailed in Section 7.7.2. The misidentification of leptons give rise to fakes and non-prompt leptons, as described in Section 7.7.3. Other backgrounds that have a minor effect on the analysis are described in Section 7.7.4.

7.7.1 $t\bar{t}$ +jets Background

The fractional contributions of the various backgrounds, illustrated in Figure 7.8, show that the analysis regions are dominated by the background arising from the production of $t\bar{t}$ +jets, containing the following different processes, $t\bar{t}$ +light, $t\bar{t}+ \geq 1b$, and $t\bar{t}+ \geq 1c$. These three processes may differ between models and their kinematics, such as jet kinematics. Therefore, they are treated independently in the analysis.

Events in the $t\bar{t}$ +jets samples are generated inclusive in jet flavor, but then classified into three non-overlapping sub-samples according to the flavor of the additional jets, $t\bar{t}$ +light-jets, $t\bar{t}+ \geq 1b$, and $t\bar{t}+ \geq 1c$, as described in Section 7.5.2. Figure 7.14 shows the relative abundance of the different $t\bar{t}$ +jets categories for the nominal $t\bar{t}$ +jets sample as well as the alternative samples. About 88.7% of the events are $t\bar{t}$ +light-jets, 8.6% are $t\bar{t}+ \geq 1c$ and about 2.7% are $t\bar{t}+ \geq 1b$. MC generators have been tuned in ATLAS [157] to agree with $t\bar{t}$ +jets measurements. As one can see from Figure 7.14, $t\bar{t}$ +light-jets processes are the dominant ones and a good description of these process is expected. While, $t\bar{t}$ +HF processes have not been measured precisely.

Since the $t\bar{t}$ +HF production cross section is not well constrained from past measurements, the normalization of the $t\bar{t}+ \geq 1b$ and $t\bar{t}+ \geq 1c$ backgrounds are determined from the fit to data, as described in Section 7.10. Therefore, the $t\bar{t}$ alternative samples, shown in Figure 7.14, are re-weighted in such a way that they have the same fractions of $t\bar{t}+ \geq 1b$, $t\bar{t}+ \geq 1c$, and $t\bar{t}$ +light as the nominal POWHEG+PYTHIA 8 sample.

The $t\bar{t}$ +HF events are further categorized using a finer classification to account for the differences in the event generators and to asses the uncertainties related to the modeling

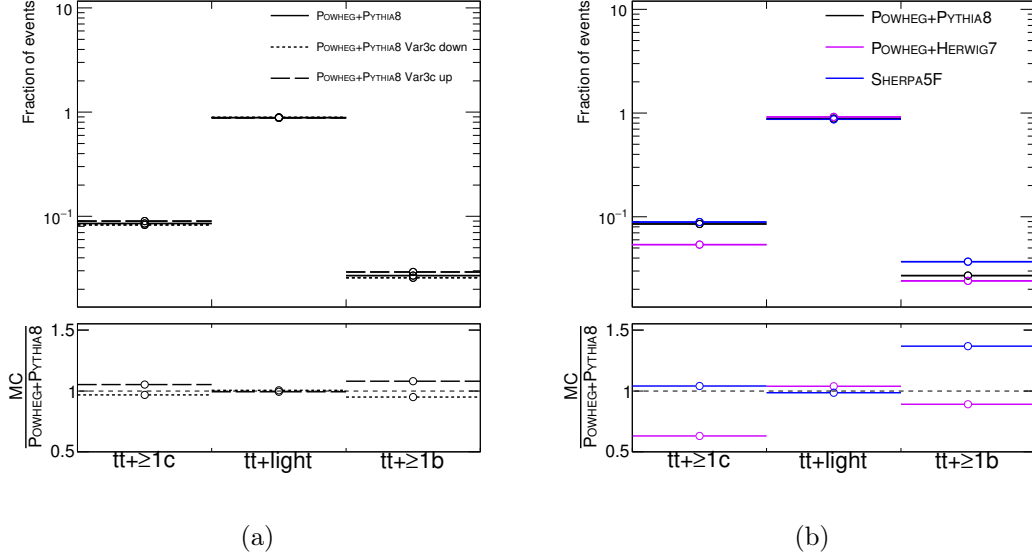


Figure 7.14: Relative abundance of $t\bar{t} + \geq 1c$, $t\bar{t} + \text{light}$, and $t\bar{t} + \geq 1b$ for the nominal POWHEG-BOX+PYTHIA 8 (solid black line), as well as the systematic samples: (a) the impact of factorization and renormalization scale variations, and the radiation systematics for POWHEG-BOX+PYTHIA 8 sample (dashed black line), (b) parton shower systematic using POWHEG+HERWIG 7 (purple line), generator systematic using SHERPA5F (blue line). Particle jets are required to have $p_T > 15$ GeV and $|\eta| < 2.5$.

of the $t\bar{t} + HF$. This categorization depends on the number of heavy hadrons and particle jets in the event. Therefore, events with exactly two additional b -jets are labelled as $t\bar{t} + b\bar{b}$, those with only one b -jet are labelled as $t\bar{t} + b$, those where a single particle jet is matched to a b -hadron pair are labelled as $t\bar{t} + B$, and the other events containing b hadrons are labelled as $t\bar{t} + \geq 3b$. In addition to these categories, there is a small contribution of events with $b\bar{b}$ pairs arising from multi-parton interactions (MPI) overlaying $t\bar{t}$ -jets events, and the production of a $b\bar{b}$ pair from a gluon radiated off the decay products of the top-quark, labeled as final-state radiation (FSR). This category is referred to as $t\bar{t} + b(\text{MPI/FSR})$. Background events from $t\bar{t}$ containing an extra charm-jet are divided analogously. The coarser classification is used to define the background categories in the likelihood fit, while this classification is used to assign correction factors and estimate uncertainties. The correction factors referred to as re-weighting will be discussed in the following section.

7.7.2 $t\bar{t} + \geq 1b$ Background

The $t\bar{t} + \geq 1b$ is the dominant irreducible background in the signal regions. Therefore a good estimate of this background is a crucial part in the $t\bar{t}H(H \rightarrow b\bar{b})$ search. The production of $t\bar{t} + \geq 1b$ in the POWHEG generator is calculated at LO for diagrams of the type $gb \rightarrow t\bar{t}b$, and at leading-logarithmic (LL) accuracy through the parton shower for processes involving a $b\bar{b}$ pair. However, fixed-order NLO calculations for the $t\bar{t} + b\bar{b}$ process can reduce perturbative uncertainties on the cross section from 70 – 80% of the LO calculation, down to about 20 – 30% [189–191]. NLO predictions with massive b -quarks in the four-flavor number scheme (4FNS) matched to a parton shower [173] are available from SHERPA+OPENLOOPS [73, 75], referred to as SHERPA4F.

The SHERPA4F sample represents the state-of-the-art theoretical knowledge of the $t\bar{t} + b\bar{b}$ process. The presence of massive b -quarks in the matrix element allows the computation to cover the full $t\bar{t} + b\bar{b}$ phase space, without artificial cut to avoid divergences at low energy or low opening angles as done in the case where b -quarks are assumed to be massless. Hence, SHERPA4F is expected to provide a more accurate estimate than POWHEG. Therefore, events of the $t\bar{t} + \geq 1b$ background are re-weighted to the NLO prediction based on the $t\bar{t} + b\bar{b}$ SHERPA4F sample. This re-weighting is performed for the different categories of $t\bar{t} + \geq 1b$, in such a way that the relative normalization of each of the sub-categories, $t\bar{t} + b$, $t\bar{t} + b\bar{b}$, $t\bar{t} + B$, $t\bar{t} + \geq 3b$, are at NLO accuracy. This is referred to as normalization or "norm re-weighting". Figure 7.15 shows the relative abundance of different $t\bar{t} + \geq 1b$ event categories for the nominal $t\bar{t}$ +jets sample compared to the $t\bar{t} + b\bar{b}$ SHERPA4F sample. The various sub-categories show that SHERPA4F predicts a higher contribution in the categories where the production of a second $b\bar{b}$ pair is required. Re-weighting is used instead of the full simulated SHERPA4F sample in order to overcome the present challenge in merging the $t\bar{t} + b\bar{b}$ from the 4F scheme with the $t\bar{t}$ +light jets in the 5F scheme.

The alternative $t\bar{t}$ samples are also re-weighted to the NLO SHERPA4F prediction prior to evaluating the relevant uncertainty. The remaining differences are referred to as "residual" uncertainties. Figure 7.16 shows the relative abundance of the different $t\bar{t} + \geq 1b$ sub-categories for the alternative $t\bar{t}$ samples compared to the nominal POWHEG-BOX+PYTHIA 8. Differences up to 30% are observed.

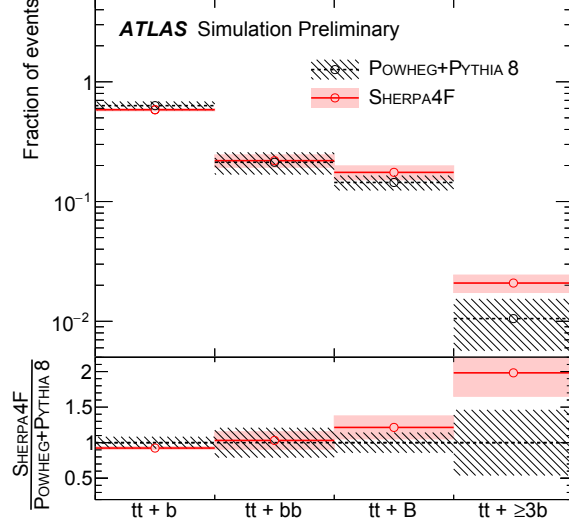


Figure 7.15: The predicted fractions for the $t\bar{t} + \geq 1b$ sub-categories. The inclusive POWHEG+PYTHIA 8 prediction, with its uncertainties arising from comparison with samples with ISR/FSR variations, different parton shower and hadronization model and different NLO generator, as illustrated in Figure 7.16 shown as the hashed area, is compared to the four-flavor $t\bar{t}$ calculation from SHERPAOL, with its uncertainties coming from a combination of various sources, as illustrated in Figure 7.17 shown as the shaded area. Particle jets are required to have $p_T > 15$ GeV and $|\eta| < 2.5$.

Additional uncertainties on the NLO prediction are considered. These are evaluated by varying the renormalization scale up and down by a factor of two, changing the functional form of the resummation scale to μ_{CMMPs} , and adopting a global scale choice, $\mu_Q = \mu_R = \mu_F = \mu_{\text{CMMPs}}$. Additionally, two alternative PDF sets, MSTW [192] and NNPDF, are considered, as well as an alternative shower recoil scheme. The variation of the SHERPA4F are shown in Figure 7.17. These uncertainties are assessed by re-weighting the $t\bar{t}$ POWHEG+PYTHIA 8 nominal sample to the SHERPA4F variations, and then taking the remaining differences.

Figure 7.18 shows Feynman diagram examples of $t\bar{t} + b\bar{b}$ -like processes, arising from MPI and FSR. These processes are estimated to be about 10% of the $t\bar{t} + \geq 1b$ events in the POWHEG+PYTHIA 8 inclusive sample, and they are not included in the NLO calculation of the 4-flavor scheme. Therefore, they are treated separately and are excluded from the re-weighting to the SHERPA4F sample.

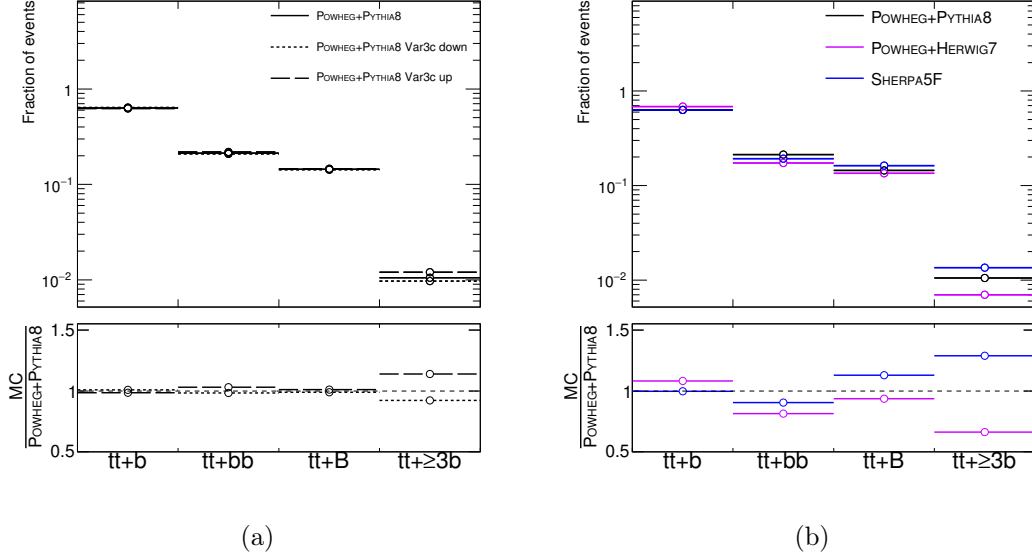


Figure 7.16: The predicted fractions for the $t\bar{t} + \geq 1b$ sub-categories. The nominal POWHEG-BOX+PYTHIA 8 (solid black line) is compared to the $t\bar{t}$ alternative samples used to assess the systematic uncertainties: (a) the impact of factorization and renormalization scale variations, and the radiation systematics for POWHEG-BOX+PYTHIA 8 sample (dashed black line), (b) parton shower systematic using POWHEG+HERWIG 7 (purple line), generator systematic using SHERPA5F (blue line). Particle jets are required to have $p_T > 15$ GeV and $|\eta| < 2.5$.

Since the relative normalization of the $t\bar{t} + \geq 1b$ events in the 5F samples are re-weighted to match the predicted fractions in SHERPA4F (norm re-weighting), only the differences in the shape of the kinematic distributions remain as systematic uncertainties between the 4 and 5 flavor schemes. Several normalized distributions of various variables in the $t\bar{t} + b\bar{b}$ category are shown in Figures 7.19, and 7.20, and in the $t\bar{t} + B$ category in Figures 7.21 and 7.22. The full set of figures can be found in Appendix A.3. The distributions of the top-quark p_T and $t\bar{t}$ p_T in the $t\bar{t} + b\bar{b}$ (Figure 7.19) and $t\bar{t} + B$ (7.21) categories show a reasonable agreement between the POWHEG+PYTHIA 8 and SHERPA4F samples. Some differences between the 5F and the 4F scheme are observed in the transverse momentum of the two additional b -jets (p_T^{bb}), and the opening angle between the two additional (b -jets ΔR^{bb}) in Figure 7.20, and the transverse momentum of the additional b -jet (p_T^b) in Figure 7.22. These differences occur from the differences in the production of $b\bar{b}$ pairs which originates only from the parton shower in the 5F scheme, compared to the 4F scheme where the full $pp \rightarrow t\bar{t}b\bar{b}$ process is in the matrix element.

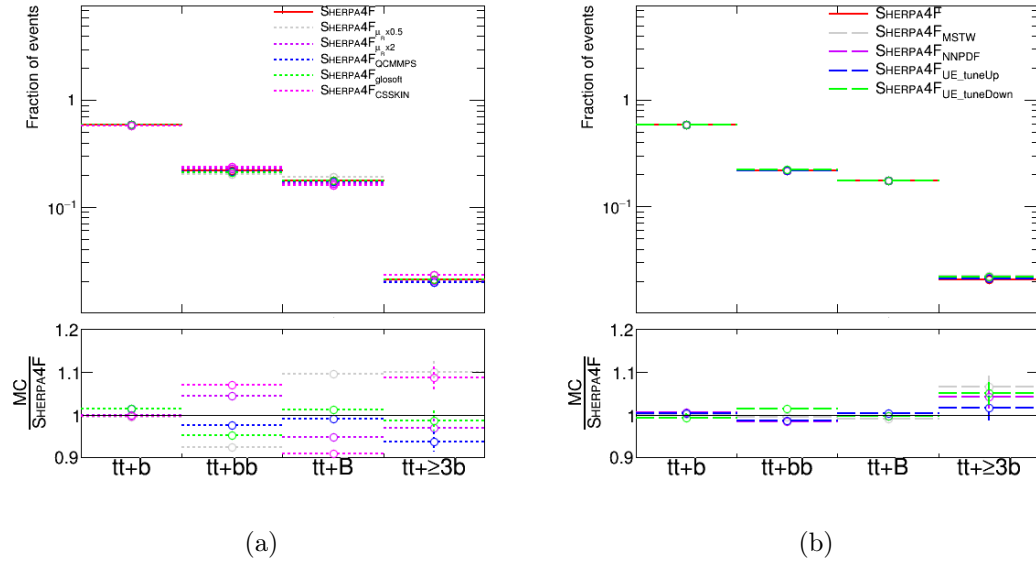


Figure 7.17: Effect of the scale variations, PDF variations, shower recoil scheme and underlying event tune on the relative contributions across the $t\bar{t} + \geq 1b$ categories. Particle jets are required to have $p_T > 15$ GeV and $|\eta| < 2.5$.

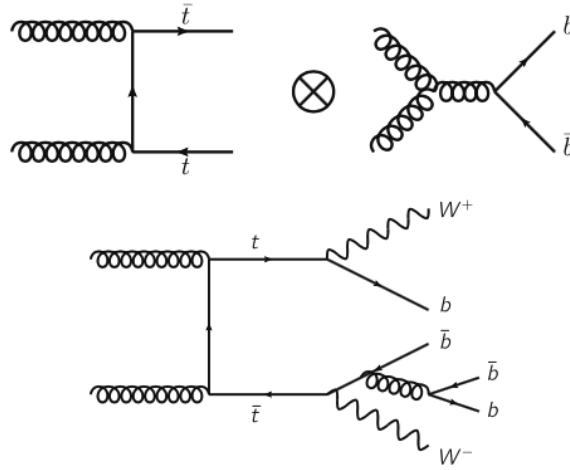


Figure 7.18: $b\bar{b}$ production from multiple parton interaction (MPI) overlaid with a $t\bar{t}$ events from the hard scatter (top diagram) and final state radiation (FSR) (bottom diagram).

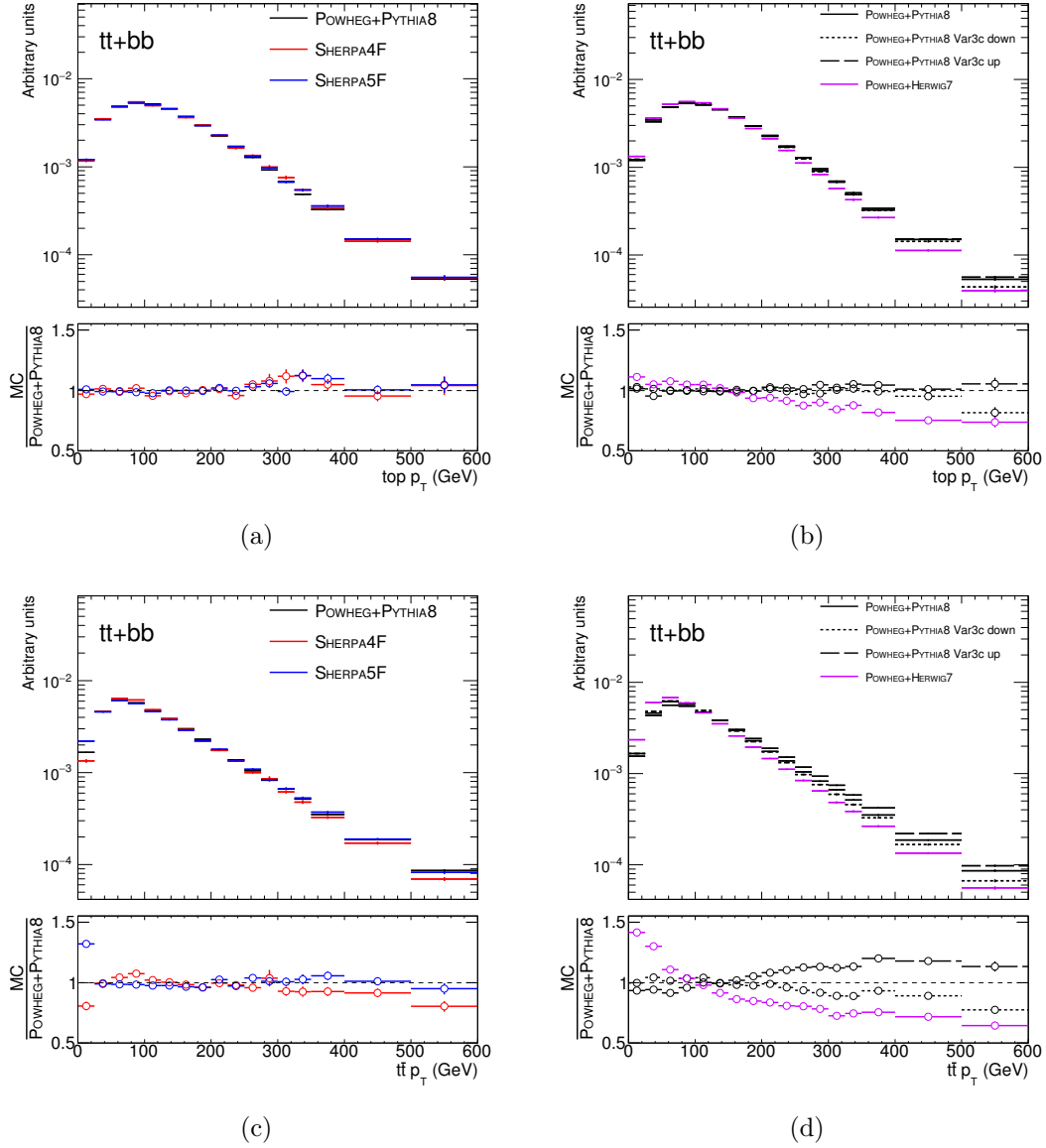


Figure 7.19: Comparison of normalized kinematic variables in the $t\bar{t} + b\bar{b}$ category: (a) and (b) show top-quark transverse momentum (p_T^{top}), (c) and (d) show the transverse momentum of the $t\bar{t}$ system, ($p_T^{t\bar{t}}$). (a) and (c) show the differences among the 5F and 4F scheme by comparing POWHEG+PYTHIA 8 and SHERPA4F. (b) and (d) show the differences among the nominal POWHEG+PYTHIA 8 and the $t\bar{t}$ alternative samples: the impact of factorization and renormalization scale variations, and the radiation systematics for POWHEG-BOX+PYTHIA 8 sample (dashed black line), parton shower systematic using POWHEG+HERWIG 7 (purple line), and generator systematic using SHERPA5F (blue line). Particle jets are required to have $p_T > 15$ GeV and $|\eta| < 2.5$.

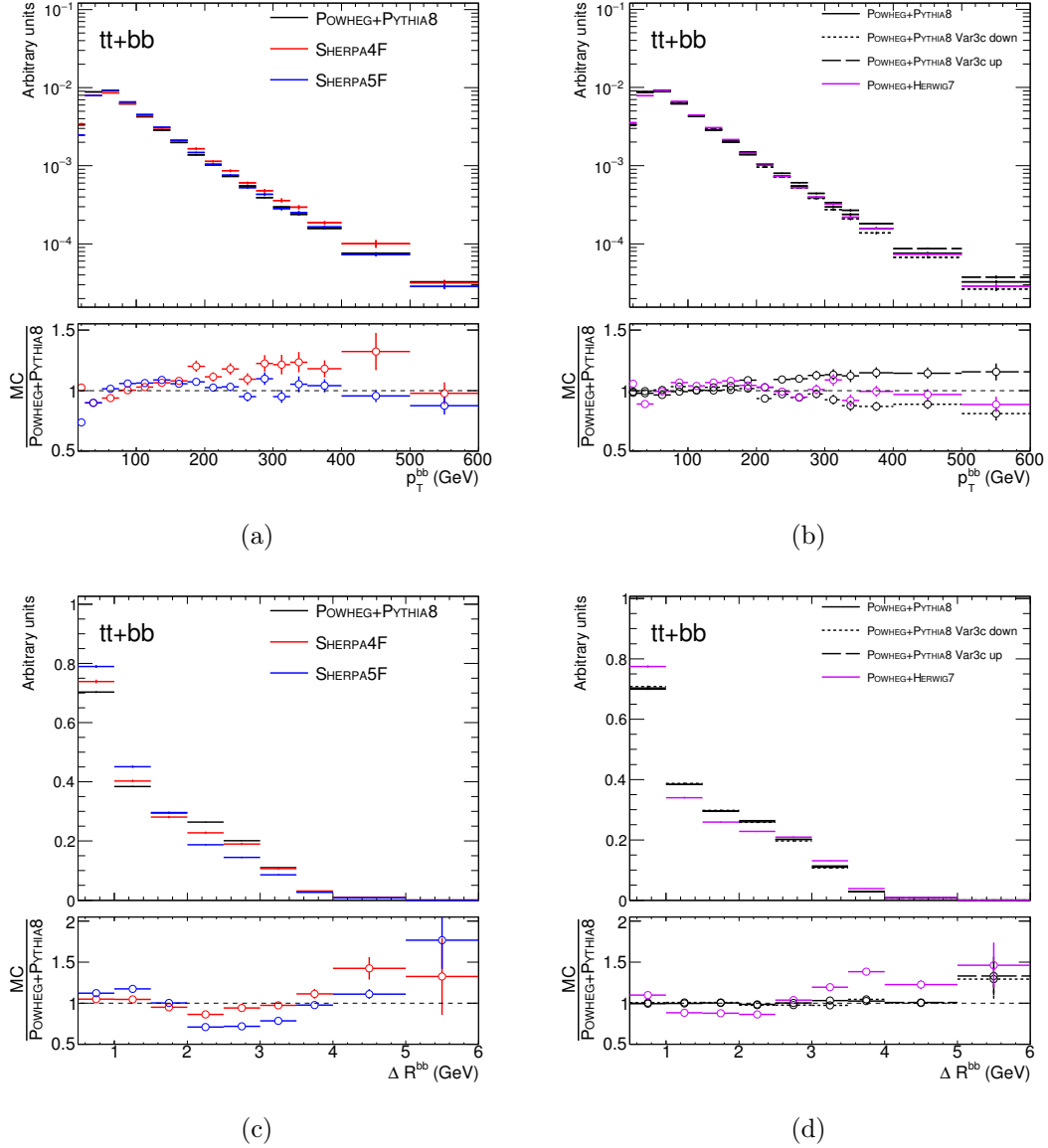


Figure 7.20: Comparison of normalized kinematic variables in the $t\bar{t} + b\bar{b}$ category: (a) and (b) show the transverse momentum of the two additional b -jets (p_T^{bb}) that do not originate from the decay of the $t\bar{t}$ system, (c) and (d) show the opening angle between the two additional (b -jets ΔR^{bb}). (a) and (c) show the differences among the 5F and 4F scheme by comparing POWHEG+PYTHIA 8 and SHERPA4F. (b) and (d) show the differences among the nominal POWHEG+PYTHIA 8 and the $t\bar{t}$ alternative samples: the impact of factorization and renormalization scale variations, and the radiation systematics for POWHEG-BOX+PYTHIA 8 sample (dashed black line), parton shower systematic using POWHEG+HERWIG 7 (purple line), and generator systematic using SHERPA5F (blue line). Particle jets are required to have $p_T > 15$ GeV and $|\eta| < 2.5$.

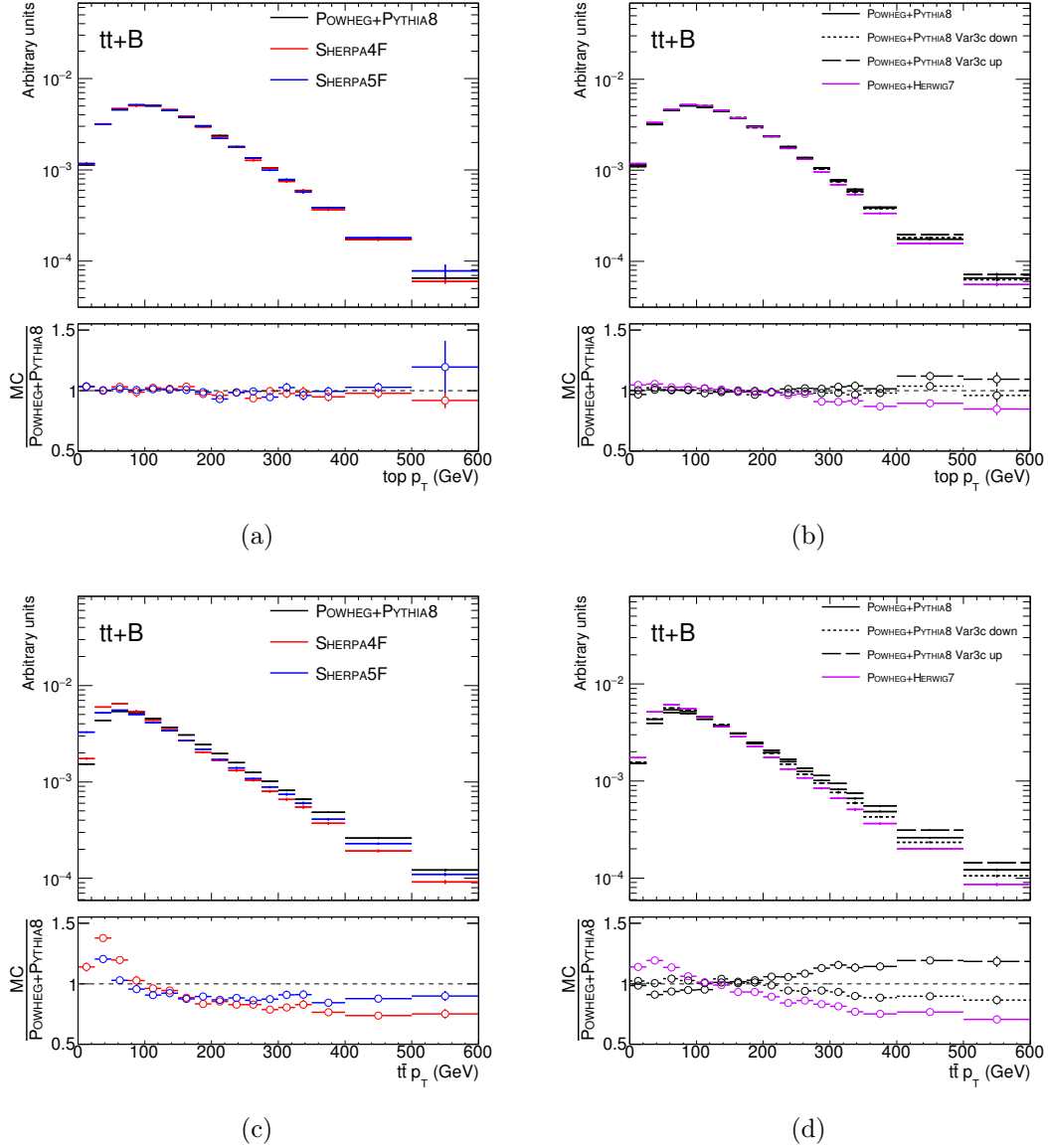


Figure 7.21: Comparison of normalized kinematic variables in the $t\bar{t} + B$ category: (a) and (b) show top-quark transverse momentum (p_T^{top}), (c) and (d) show the transverse momentum of the $t\bar{t}$ system, ($p_T^{t\bar{t}}$). (a) and (c) show the differences among the 5F and 4F scheme by comparing POWHEG+PYTHIA 8 and SHERPA4F. (b) and (d) show the differences among the nominal POWHEG+PYTHIA 8 and the $t\bar{t}$ alternative samples: the impact of factorization and renormalization scale variations, and the radiation systematics for POWHEG-BOX+PYTHIA 8 sample (dashed black line), parton shower systematic using POWHEG+HERWIG 7 (purple line), and generator systematic using SHERPA5F (blue line). Particle jets are required to have $p_T > 15$ GeV and $|\eta| < 2.5$.

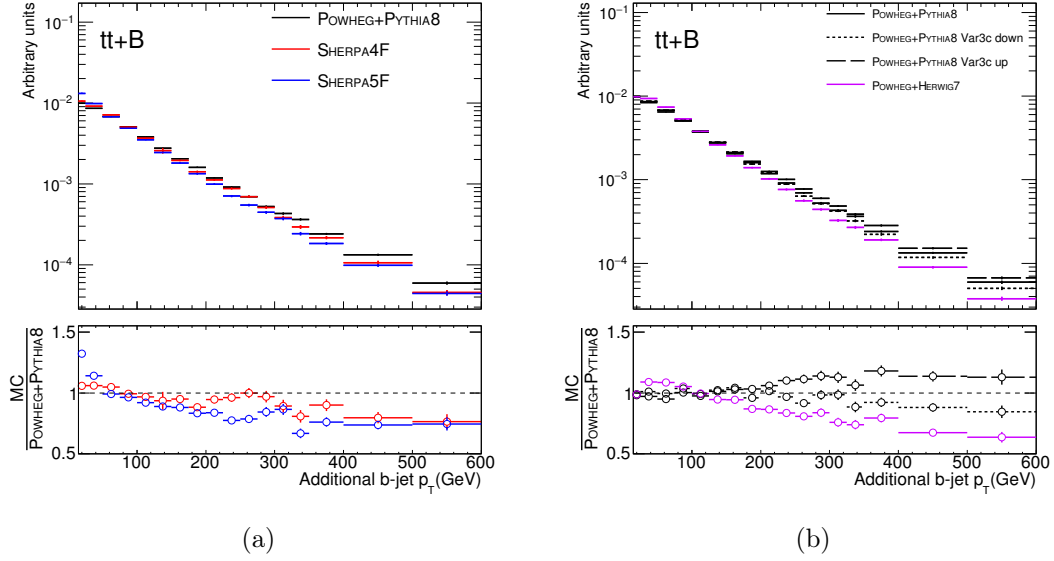


Figure 7.22: Comparison of the normalized transverse momentum of the additional b -jet (p_T^b) that does not originate from the decay of the $t\bar{t}$ system, in the $t\bar{t} + B$ category. (a) shows the differences among the 5F and 4F scheme by comparing POWHEG+PYTHIA 8 and SHERPA4F. (b) shows the differences among the nominal POWHEG+PYTHIA 8 and the $t\bar{t}$ alternative samples: the impact of factorization and renormalization scale variations, and the radiation systematics for POWHEG-BOX+PYTHIA 8 sample (dashed black line), parton shower systematic using POWHEG+HERWIG 7 (purple line), and generator systematic using SHERPA5F (blue line). Particle jets are required to have $p_T > 15$ GeV and $|\eta| < 2.5$.

Given the differences between the predictions of POWHEG-BOX+PYTHIA 8 and SHERPA4F a re-weighting procedure of the differential distributions was tried to improve the modeling. This so called *shape re-weighting* consists of sequential re-weightings based on one-dimensional distributions related to the kinematics of the top-quark and the additional b -jets. The shape re-weighting is first based on the p_T of the $t\bar{t}$ system, followed by a sequential re-weighting to the top-quark p_T . A third re-weighting was then chosen depending on the type of the event under consideration. In topologies with only one additional b -jet, the p_T of that jet was used as a last re-weighting and in topologies with more than one additional b -jet, the ΔR between the additional b -jets was used and then the p_T of the dijet system was used as a final step of the re-weighting. Figures 7.19 and 7.20 show the comparison between the POWHEG+PYTHIA 8 and the SHERPA4F for the variables that were used to derive the shape re-weighting in the $t\bar{t} + b\bar{b}$ category. Similarly, Figures 7.21 and 7.22 show the used variables in the $t\bar{t} + B$ category.

The shape re-weighting method is validated by comparing the kinematic variables at reconstruction level with and without the re-weighting, as shown in Figures 7.23 and 7.24. The solid black line represents the nominal POWHEG+PYTHIA 8 sample, the dashed black line represents POWHEG+PYTHIA 8 sample after applying the shape re-weighting to SHERPA4F, and the red line represents the SHERPA4F. The shape re-weighting was found to have a non-significant effect on reconstructed distributions. Therefore, only the norm re-weighting is kept and the differences between the predictions from POWHEG+PYTHIA 8 and SHERPA4F in the differential distributions are considered as one additional source of uncertainty. This uncertainty does not affect the relative fractions of the $t\bar{t} + b$, $t\bar{t} + b\bar{b}$, $t\bar{t} + B$, and $t\bar{t} + \geq 3b$ sub-components as these fractions are fixed to the prediction of SHERPA4F.

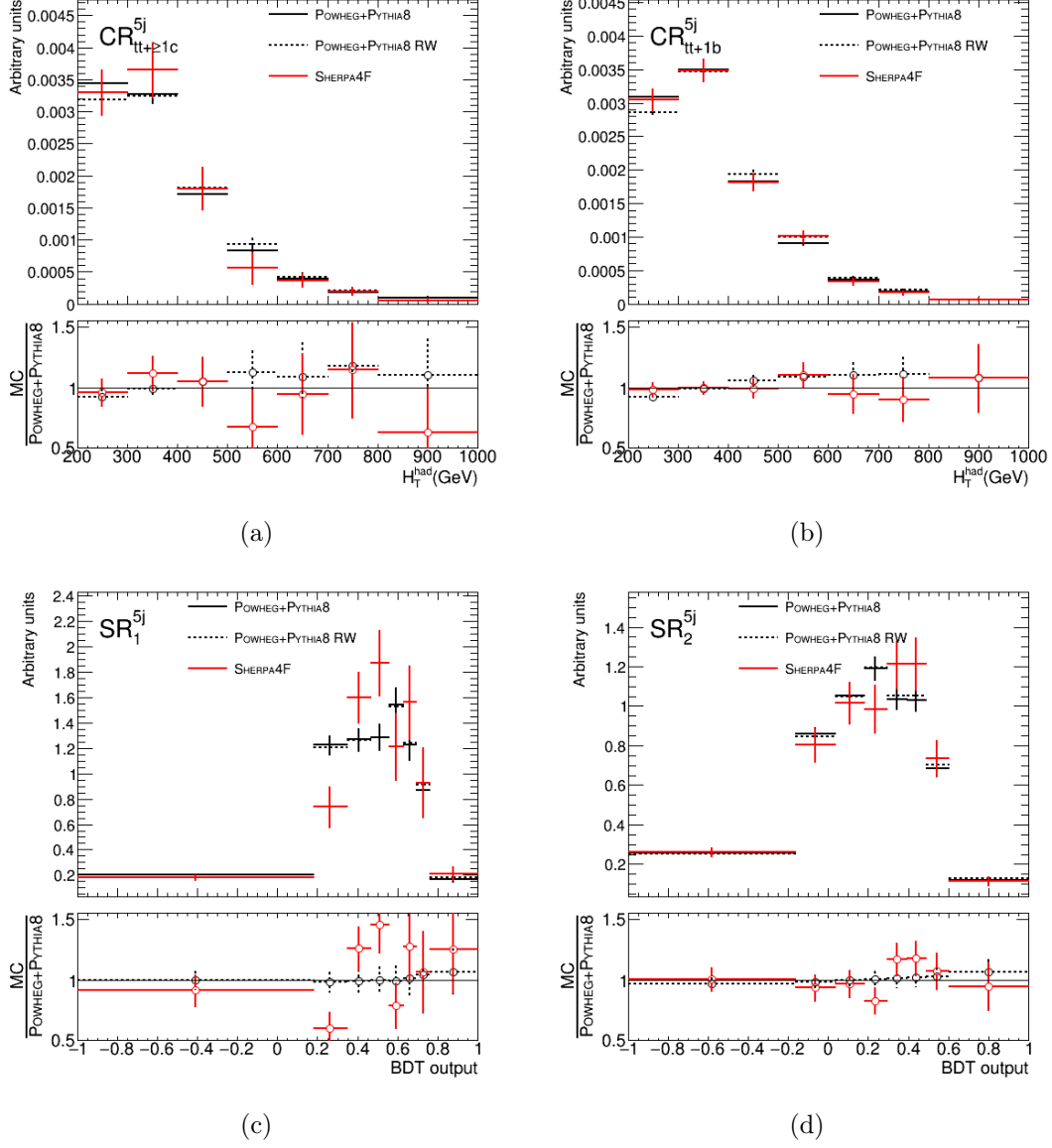


Figure 7.23: Comparisons of the normalized (a-b) H_T^{had} and (c-d) BDT output distributions in the five-jet regions in the single lepton channel. The SHERPA4F sample used to derive the re-weighting, is in red, the nominal POWHEG+PYTHIA 8 is in solid black, and POWHEG+PYTHIA 8 after the shape re-weighting (RW) to SHERPA4F is in dashed black.

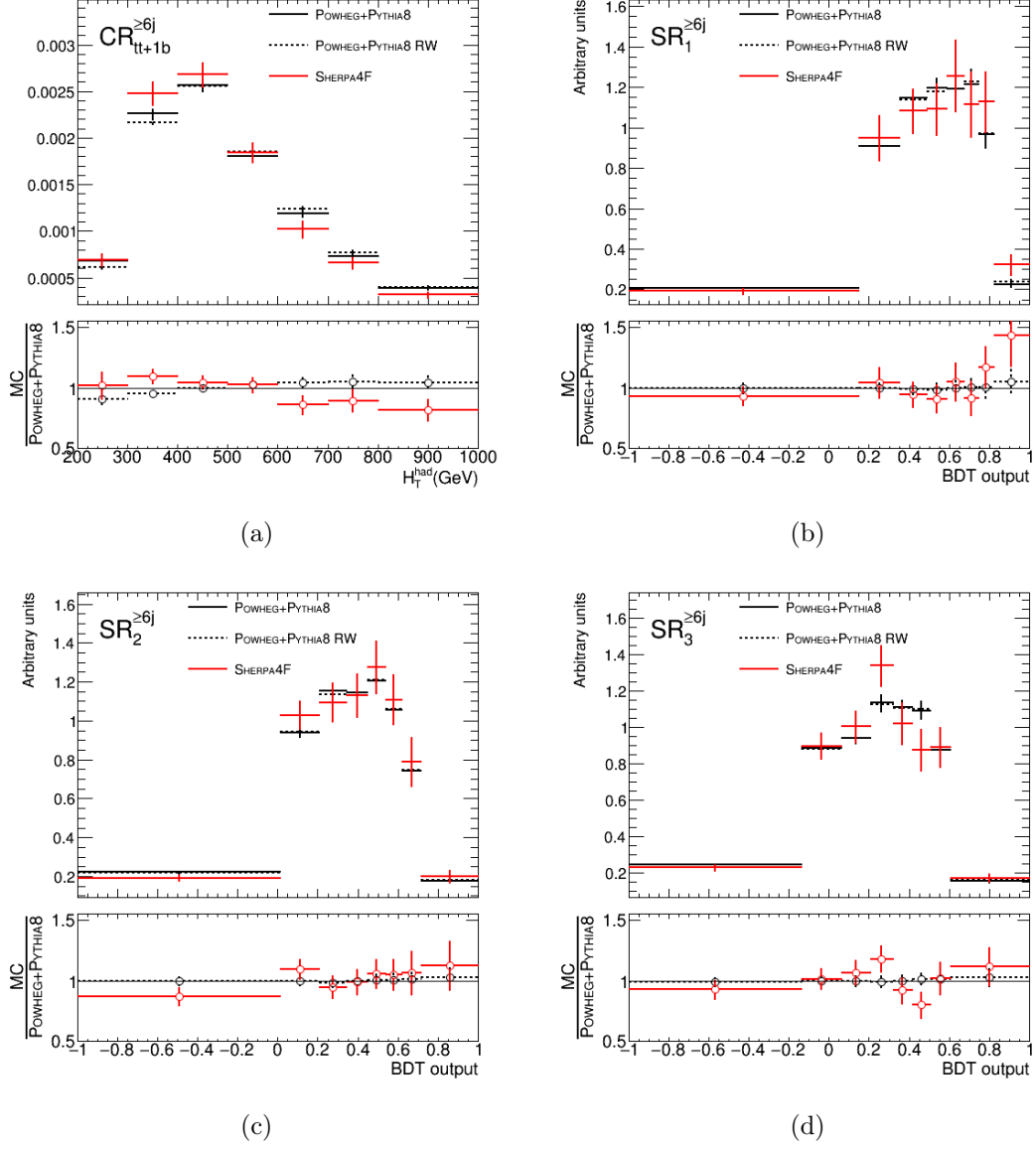


Figure 7.24: Comparisons of the normalized (a) H_T^{had} and (b-d) BDT output distributions in the six-jet regions in the single lepton channel. The SHERPA4F sample used to derive the re-weighting, is in red, the nominal POWHEG+PYTHIA 8 is in solid black, and POWHEG+PYTHIA 8 after the shape re-weighting (RW) to SHERPA4F is in dashed black.

7.7.3 Fake and Non-prompt Lepton Background

Events containing non-prompt leptons or jets misidentified as leptons at reconstruction level may satisfy the analysis selection criteria, giving rise to a background in the $t\bar{t}H(H \rightarrow b\bar{b})$ analysis regions. Such background events could arise from the fully hadronic decay of a pair of top-quarks or QCD multijet processes with many b -jets. Even though these events have small acceptance rates, their production rates are significantly larger than the processes of interest. The misidentified lepton background in the electron channel, referred to as (e+jets), contributes via the misidentification of a jet or a photon as an electron or the presence of a non-prompt electron arising from heavy-hadron decays. While in the muon channel, referred to as (μ +jets), the contribution is mainly due to a non-prompt muon arising from b - or c - hadron decays. In the dilepton channel, this background is estimated from simulation and is normalized to data in control regions with two same-sign leptons. However, in the single-lepton channel, this background is estimated using a data-driven technique based on the Matrix Method, detailed in Chapter 6. The following details the results obtained in the single-lepton channel.

The Matrix Method exploits the difference in lepton identification and isolation requirements between real, prompt leptons, and fake, non-prompt leptons. Moreover, the fake estimate is extrapolated from dedicated regions, with looser lepton identification and isolation, to the signal regions. Events are first selected using the single-lepton triggers listed in Table 7.1, then they are divided in two samples: a "Tight sample" that has events with one tight lepton and a "Loose sample" that has events with one loose lepton. The Tight sample has to be a subset of the Loose sample. The tight selection, used to define the Tight sample, applies the same requirements as used in the analysis and defined in Section 7.5. While the loose selection, used to define the Loose sample, uses looser lepton identification and isolation operating points. In the loose selection, electrons are required to pass the medium likelihood-based identification with no additional requirement on the isolation, and muons are required to pass the medium quality with no requirement on the isolation. Hence, more fake leptons are expected in the Loose sample. Both the loose and tight selections are summarized in Table 7.6.

The number of fake leptons in the Tight sample is estimated from counting the number of loose events that do not pass the Tight selection and an event weight, as defined in

	Loose Selection	Tight Selection
Electron identification level	MediumLH	TightLH
Muon identification level	Medium	Medium
Lepton isolation requirement	None	Gradient

Table 7.6: Summary of the loose and tight lepton off-line selections. These requirements ensure that the Tight sample is a subset of the Loose sample.

Equation 6.7, is computed.

Fake ϵ_f and real ϵ_r efficiencies are measured in dedicated regions which are representative of the signal regions in terms of kinematics, and the composition of the fake and non-prompt lepton background. The regions are formed from events which have at least one jet, exactly one loose lepton passing the loose selection defined in Table 7.6, and satisfies the analysis triggers listed in Table 7.1. Fake and non-prompt leptons are referred to as *fake* in the remainder of this chapter.

As discussed in Section 6.3.1, the fake-enriched regions used to estimate the fake efficiency in the e +jets channel are chosen using low E_T^{miss} with different requirements on the number of jets. The different number of jets is used to account for potential fluctuations due to the sample composition. Figure 7.25 shows the E_T^{miss} and m_T^W distributions for both the loose and tight selections using the 2016 data sample. The Loose sample has a factor of two more fakes (the region between the top of the stacked simulated processes and the data in Figure 7.25 (a) and (c)) than the Tight sample. Also, Figure 7.25 shows that fakes are expected at low E_T^{miss} and m_T^W . In the case of the μ +jets channel, the fake-enriched regions are chosen using the impact parameter significance d_0^{sig} . Figure 7.26 shows the d_0^{sig} distributions for both the loose and tight selections using the 2016 data sample. The difference between the data points and the backgrounds, from simulated MC events, is expected to account for the fake lepton background.

The fake efficiency is measured in the fake-enriched regions (CR_f), defined in Table 7.7. This is measured by taking the ratio between the tight and loose events after subtracting the sum of simulated backgrounds from data, using Equation 6.6.

The real efficiency ϵ_r is measured using the tag-and-probe method applied to data for the $Z \rightarrow \mu\mu$ and $Z \rightarrow ee$. Events with a pair of same-flavour opposite-sign loose or tight leptons and at least one jet are selected. The invariant mass of the dilepton system, shown

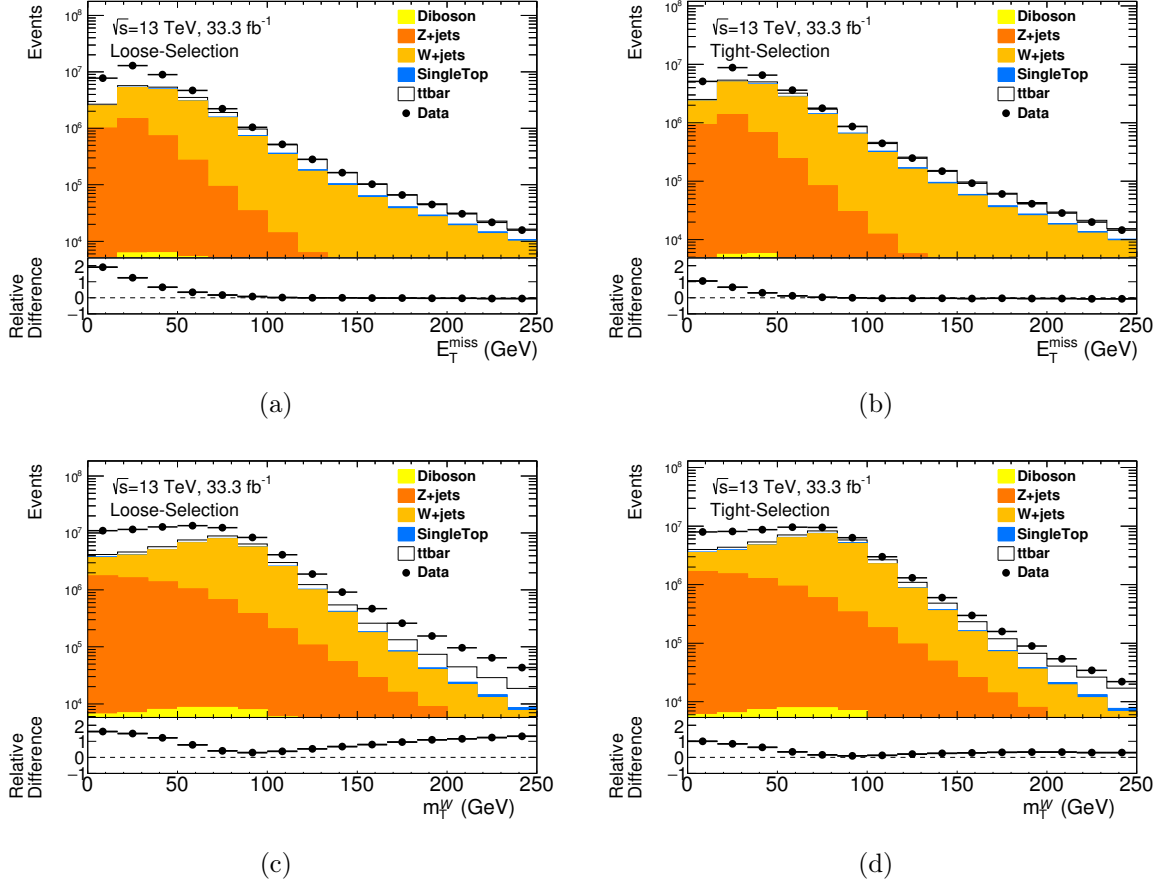


Figure 7.25: Distributions of the (a-b) E_T^{miss} and (c-d) m_T^W in e +jets events from the 2016 data, which corresponds to 33.3 fb^{-1} , and simulated MC background events. In (a) and (c), events are required to have exactly one loose electron and at least two jets, with no requests on the number of b -tagged jets. While, in (b) and (d) events are required to have exactly one tight electron. The region between the top of the stacked simulated processes and the data is assumed to come from the fake electron background contribution. The relative difference is calculated as $(\text{Data}-\text{MC})/\text{MC}$.

in Figure 7.27, is required to be close to the Z mass peak, between 60 and 120 GeV.

The selected $Z \rightarrow l\bar{l}$ sample might still contain about 5% percent of fake lepton backgrounds. Even after requiring the tag-and-probe pair to have opposite-sign (OS) charges, most of the background contribution arises from random combinations of two particles which do not originate from a resonance decay. However, the invariant mass of the tag-and probe pair, as shown in Figure 7.27, is used to discriminate signal leptons against background. This background is determined using a side band subtraction approach where

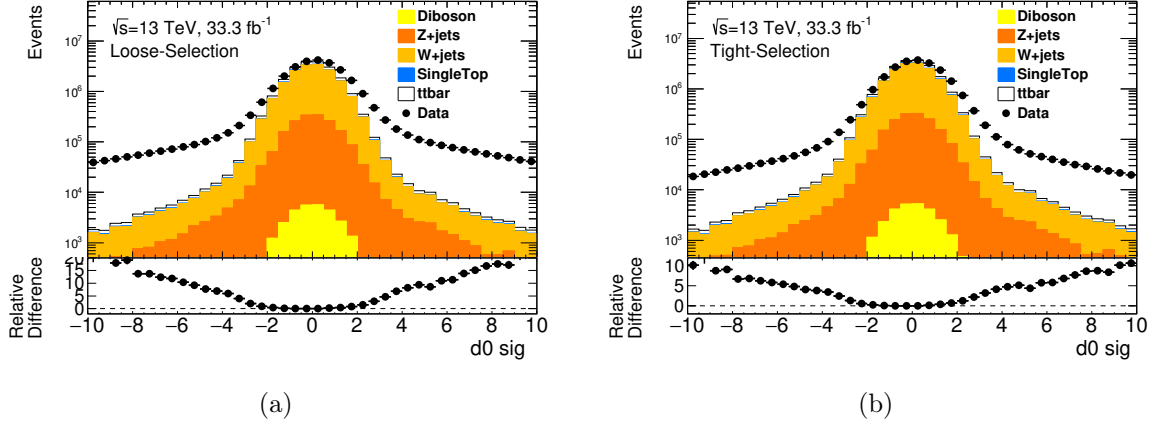


Figure 7.26: Distributions of the transverse impact parameter significance d_0^{sig} in μ +jets events from the 2016 data, which corresponds to 33.3 fb^{-1} , and simulated MC background events. In (a), events are required to have exactly one loose muon and at least two jets, with no requests on the number of b-tagged jets. While, in (b) events are required to have exactly one tight muon. The region between the top of the stacked simulated processes and the data is assumed to come from the fake muon background contribution. The relative difference is calculated as $(\text{Data-MC})/\text{MC}$.

Channel	n_{jet}	Selection	Other Selection
e+jets	$= 1$	jet	$E_T^{miss} < 20 \text{ GeV}$
e+jets	≥ 2	jets	$E_T^{miss} < 20 \text{ GeV}$
μ +jets	$= 1$	jet	$ d_0^{sig} > 5$
μ +jets	≥ 1	jet	$ d_0^{sig} > 5$

Table 7.7: Summary of the various fake-enriched regions, represented by "Other Selection", used to extract the Matrix Method fake efficiencies ϵ_f .

the real efficiency ϵ_r is extracted in each considered bin in the distribution. This method relies on the background having a smoothly falling shape over the considered invariant mass range. The invariant mass distributions for opposite-sign (black in Figure 7.27) and same-sign (grey in Figure 7.27) pairs are divided in three regions: $A = [61 - 81] \text{ GeV}$, $B = [81 - 101] \text{ GeV}$, and $C = [101 - 121] \text{ GeV}$.

Both the real ϵ_r and fake ϵ_f efficiency are measured as a function of several observables such as lepton p_T and η , leading jet p_T , the angular distance between the lepton and the closest jet $\min \Delta R(l, \text{jet})$, and the transverse plane between the lepton and the $E_T^{miss}(\Delta \Phi(l, \text{MET}))$ in one dimension parameterized with respect to the number of jets (1

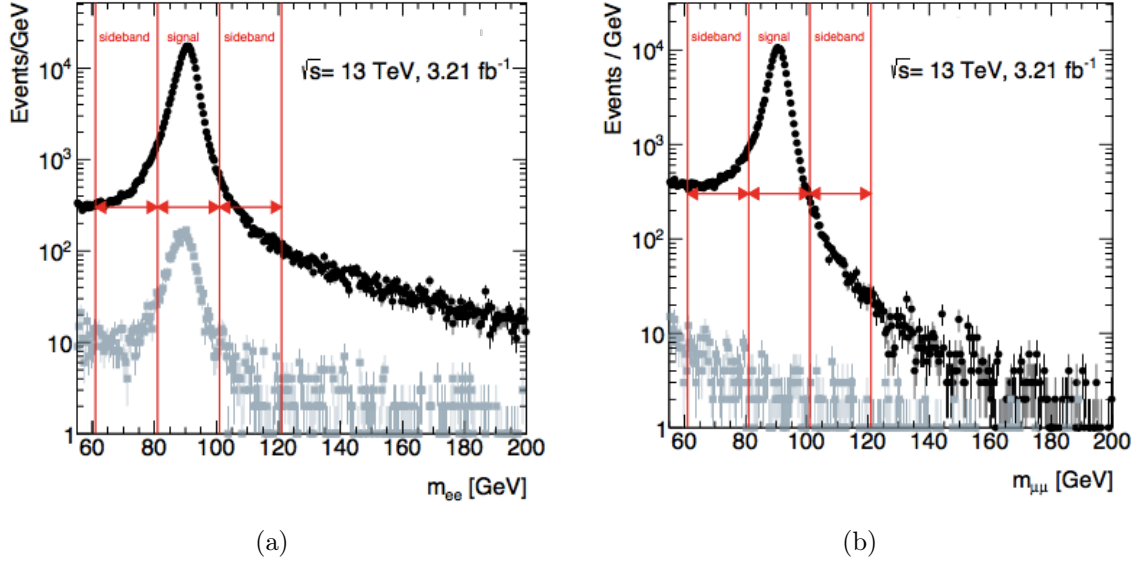


Figure 7.27: Distribution of the invariant mass of opposite (black) and same (grey) sign charge loose (a) electron and (b) muon pairs. Lines show the signal and sideband regions where the yields are calculated.

jet exclusive and ≥ 2 jets). The real ϵ_r and fake ϵ_f efficiencies are shown in Figure 7.28 for e +jets channel and in Figure 7.29 for μ +jets channel using the 2016 data and requiring events to have at least two jets.

The real efficiency (in red in Figure 7.28) is about 90% and is compatible with the electron identification efficiency using tag-and-probe in Figure 5.3. The real efficiency distributions only show slight dependence on the kinematic variables used for the parametrization and flat in most of the cases. However, the fake efficiency distributions vary between 40% and 60% in the case of e +jets, and between 20% and 60% in the case of μ +jets. In particular, the fake efficiencies show a strong dependence on the lepton p_T (Figure 7.28 (a) and Figure 7.29 (a)). More fake events are expected at low- p_T compared to high- p_T , as shown in Figure 7.28 (a) and Figure 7.29 (a), where the fake efficiency varies up to 60% at high- p_T . The discontinuity in the spectrum is affected by the turn-on of the trigger at 60 GeV for electrons and 50 GeV for muons.

The fake efficiency as a function of η varies up to 10% in the case of e +jets (Figure 7.28 (b)), and up to 40% in the case of μ +jets (Figure 7.29 (b)). A dip in the fake efficiency is noticed in the case of e +jets, which is due to the transition region between the barrel and

the end-cap EM calorimeter ($1.34 < |\eta| < 1.52$).

The fake efficiency has a small dependence on the p_T of the leading jet. A variation of about 10% is shown in Figure 7.28 (c) and Figure 7.29 (c). Similarly, the angular distance between the lepton and the closest jet $\min\Delta R(l, \text{jet})$ shows a 20% variation in the case of e +jets (Figure 7.28 (d)) and 10% in the case of μ +jets (Figure 7.28 (d)). On the other hand, a larger variation is observed in the $\Delta\Phi(l, \text{MET})$ distribution. A difference of about 25% in the case of e +jets (Figure 7.28 (e)) and about 60% in the case of μ +jets (Figure 7.29 (e)). The difference seen among the electrons and muons arises from the difference in the source of the fakes in each channel. In the muon channel, fakes are most likely to be caused by b - or c -hadron decays, whereas in the electron channel, fakes are mainly due to a misidentification of a jet. In the case of a fake muon, $\Delta\Phi$ between the lepton and the E_T^{miss} is smaller in regions with more fake muons, as shown in Figure 7.29 (e). In the case of electrons, the direction of the E_T^{miss} would be away from a fake electron which was misidentified as a jet. Moreover, the region to estimate the fake efficiency is chosen by requiring $E_T^{\text{miss}} < 20$ GeV, which is not the case for muons. All these differences result in a different dependence on $\Delta\Phi(l, \text{MET})$ among the electron and muon channels.

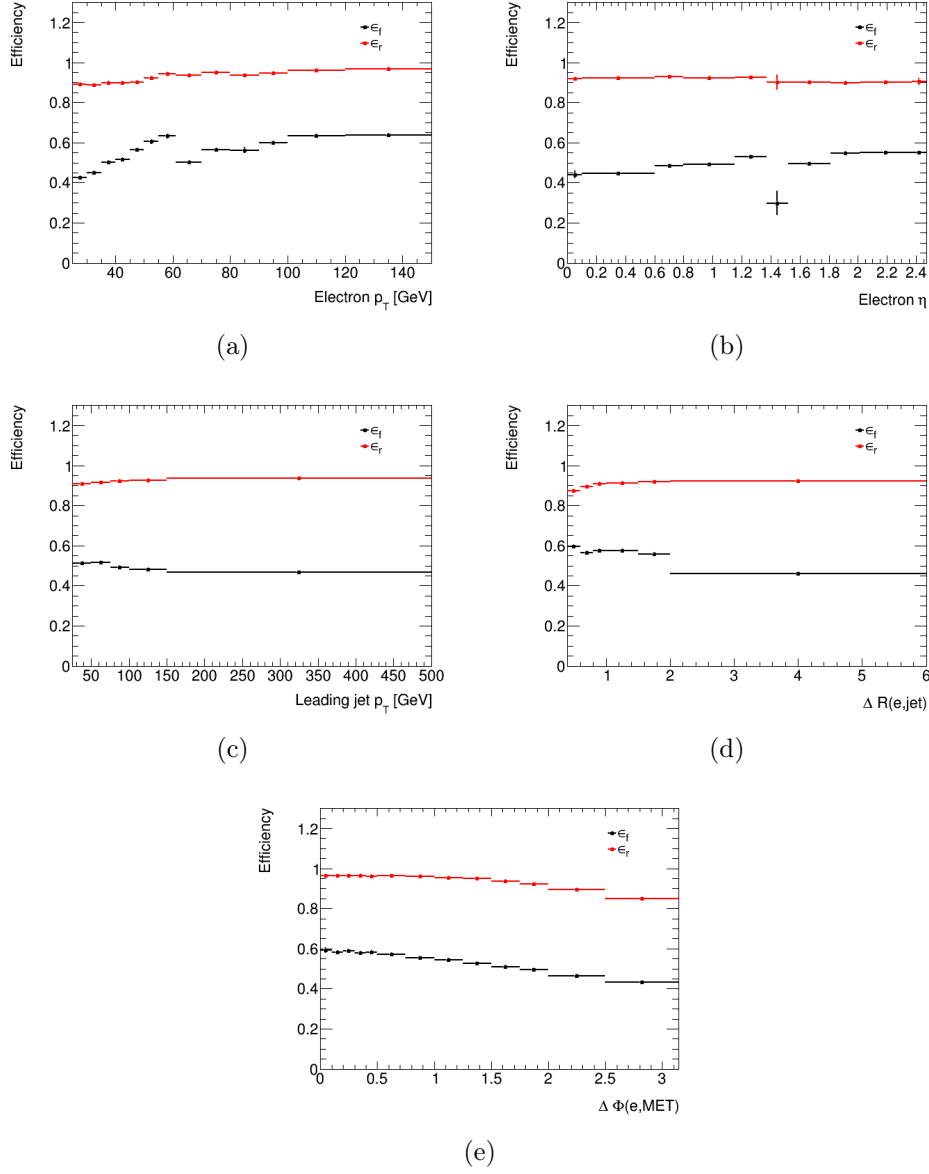


Figure 7.28: Real ϵ_r (in red) and fake ϵ_f (in black) efficiencies as measured in the 2016 data in the e +jets channel, requiring at least two jets, as function of (a) the electron p_T , (b) the electron η , (c) leading jet p_T , (d) the angular distance between the electron and the closest jet $\min \Delta R(e, \text{jet})$, and (e) the transverse plane between the electron and the $E_T^{\text{miss}}(\Delta \Phi(e, \text{MET}))$.

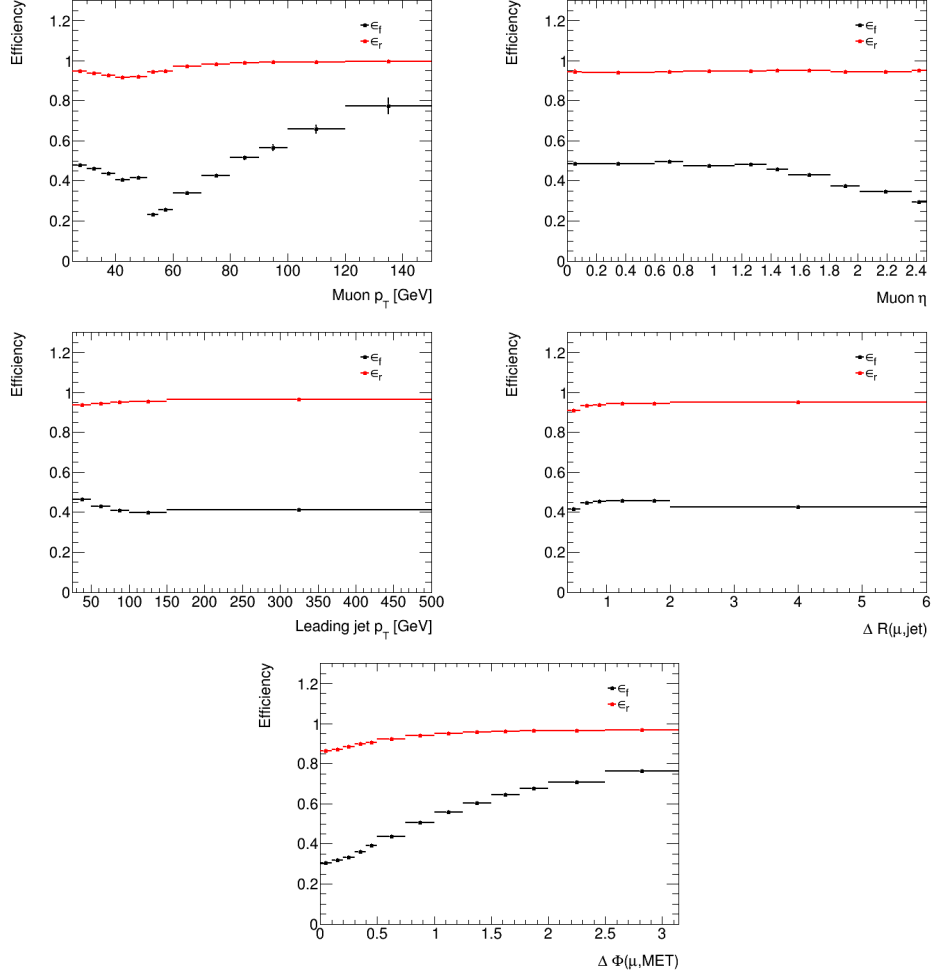
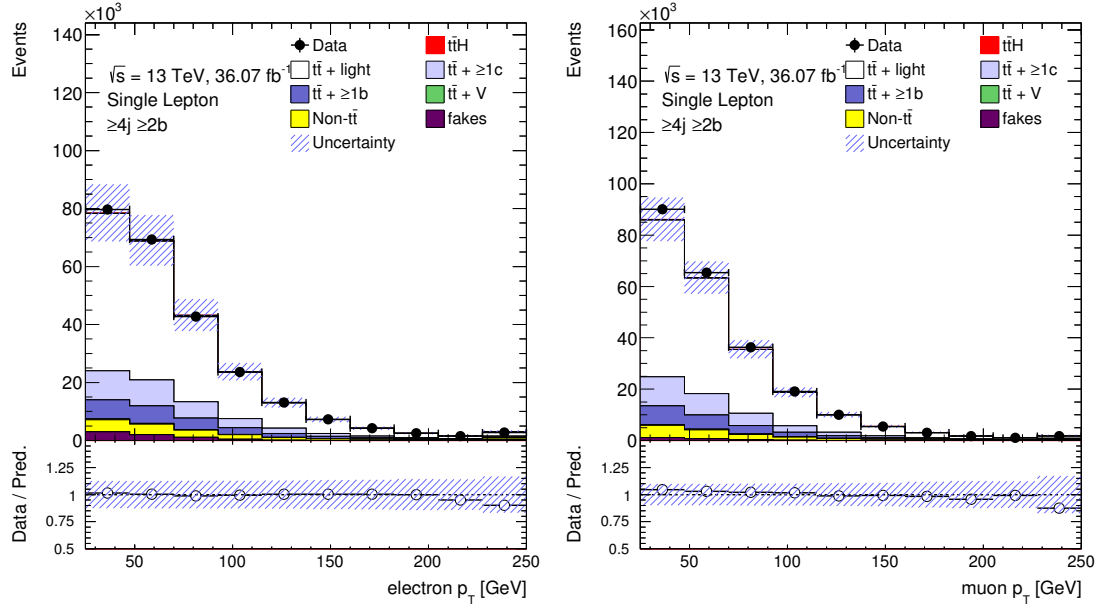


Figure 7.29: Real ϵ_r (in red) and fake ϵ_f (in black) efficiencies as measured in the 2016 data in the μ +jets channel, requiring at least two jets, as function of (a) the muon p_T , (b) the muon η , (c) leading jet p_T , (d) the angular distance between the muon and the closest jet $\min \Delta R(\mu, \text{jet})$, and (e) the transverse plane between the electron and the $E_T^{\text{miss}}(\Delta \Phi(\mu, \text{MET}))$.

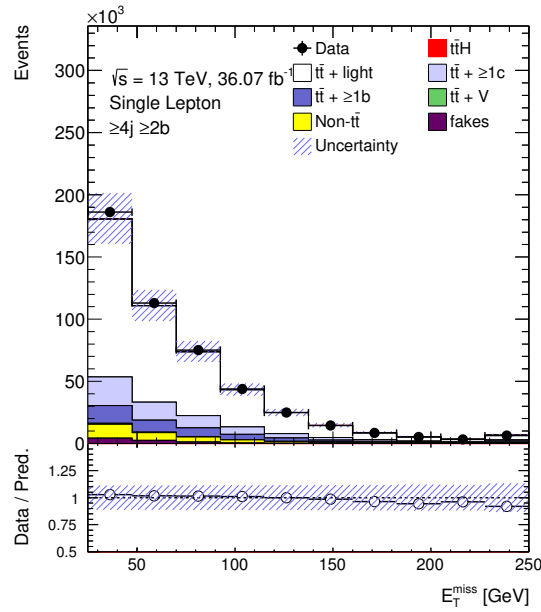
The event weight in Equation 6.7, is obtained by parametrizing the real and fake efficiencies as a function of several object kinematics, as described in Equation 6.11. The x variables in Equation 6.11 are the number of jets (one jet, or at least two jets), and the y variables are the lepton η , leading jet p_{T} , and the ΔR between the lepton and its nearest jet.

The fake estimate is validated in a region that is not included in the analysis, referred to as *validation region*, by requiring events with at least four jets, two of which are b -tagged using the 70% b -tagging operating point. Figure 7.30 shows the electron and muon p_{T} , and $E_{\text{T}}^{\text{miss}}$ in the *validation region*. Fakes are indicated by purple and not included in the non- $t\bar{t}$ background represented in yellow. More background from fake leptons is observed in the e +jets channel compared to the μ +jets channel. A small discrepancy is visible in the low- p_{T} bins, but overall a good agreement between data and prediction is observed.



(a)

(b)



(c)

Figure 7.30: Comparison of the predicted and observed p_T of the (a) electron and (muon), and the (c) E_T^{miss} distributions of the four-jet validation region in the single-lepton channel. The hashed area represent the sum of the statistical and systematic uncertainties. Distributions are shown before the fit procedure. Backgrounds from non- $t\bar{t}$ processes, excluding fakes which are represented in purple, are grouped together and represented in yellow.

Tables 7.8 and 7.9 summarize the estimate of fake lepton background in the control regions. About 1% of fake leptons are expected in the $t\bar{t}$ +light control regions. The reported errors contain only the statistical uncertainty on the estimate.

Estimate	$CR_{t\bar{t}+light}^{5j}$	$CR_{t\bar{t}+\geq 1c}^{5j}$	$CR_{t\bar{t}+1b}^{5j}$
Nominal triggers	3600 ± 200	110 ± 30	220 ± 40
Non fake background estimate	$253,000 \pm 29,000$	4000 ± 1000	9800 ± 1400

Table 7.8: Summary of the fake lepton and non-fake background estimate in the five-jet control regions, for the combined electron and muon channels. Only the statistical uncertainty is reported here.

Estimate	$CR_{t\bar{t}+light}^{\geq 6j}$	$CR_{t\bar{t}+\geq 1c}^{\geq 6j}$	$CR_{t\bar{t}+1b}^{\geq 6j}$
Nominal triggers	2000 ± 150	220 ± 40	230 ± 40
Non fake background estimate	$180,000 \pm 39,999$	9400 ± 2700	8000 ± 1200

Table 7.9: Summary of the fake lepton and non-fake background estimate in the six-jet control regions, for the combined electron and muon channels. Only the statistical uncertainty is reported here.

Studies have shown that the fake rate drops for high b -jet multiplicities, and no fake events are expected in the signal regions with four b -tagged jets at 60%. This was confirmed with the matrix method, which gave an estimate consistent with zero with large statistical uncertainties. The fake estimate in the most sensitive regions of the analysis is shown in Figure 7.31. The number of bins illustrated matches that used in the final analysis. Table 7.10 lists the yields of the fake events compared to the total number of background events in the most sensitive signal regions. The yields and the error bars in Figure 7.31 reflect the large statistical fluctuations in these regions. Note that the positive and negative yields in some regions of the phase space nearly cancel while they still have measurable errors, resulting in large statistical errors on the estimate. Therefore, fake leptons are neglected in these regions.

Estimate	$SR_1^{\geq 6J}$	$SR_2^{\geq 6J}$	SR_1^{5J}
Nominal triggers	16 ± 13	43 ± 19	3 ± 7
Non-fake background estimate	1200 ± 240	2300 ± 400	370 ± 70

Table 7.10: Summary of the fake lepton and non-fake background estimate in the most sensitive signal regions for the combined electron and muon channels.

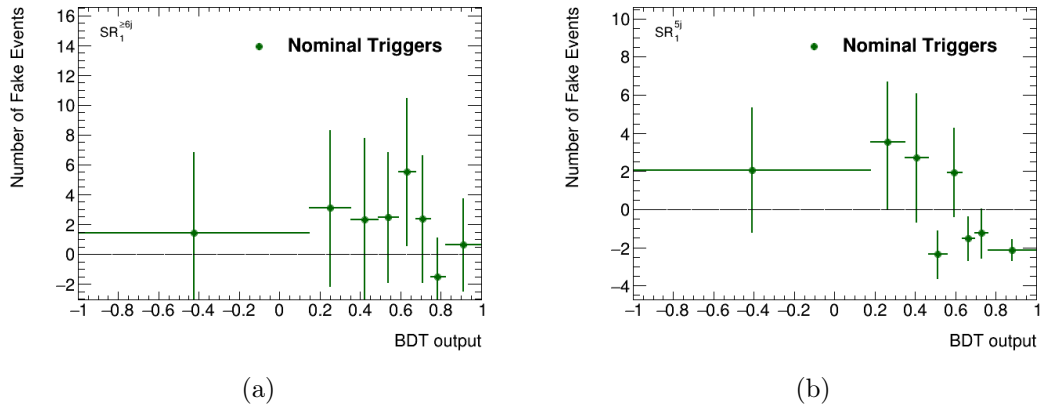


Figure 7.31: Distributions of the BDT output in (a) $SR_1^{>6J}$, and (b) SR_1^{5J} , showing the estimate of the fake background obtained from the matrix method. The error bars represent the statistical uncertainty.

Cross-checks of the Fake Estimate

1- The Choice of Parametrization

Various combinations of the parameterizations, shown in Figure 7.28 and Figure 7.29, were studied and gave compatible fake estimate within 10%. An example is shown in Figure 7.32, where the fake estimate in *option A* (in black in Figure 7.32) is parametrized as function of lepton p_T , $\min\Delta R(l, \text{jet})$, and lepton η , and compared to the nominal estimate in *option B* (in red in Figure 7.32) which is parametrized as function of leading jet p_T , $\min\Delta R(l, \text{jet})$, and lepton η . The difference between the various options is considered as part of the assigned systematic uncertainties on the estimate of fake leptons.

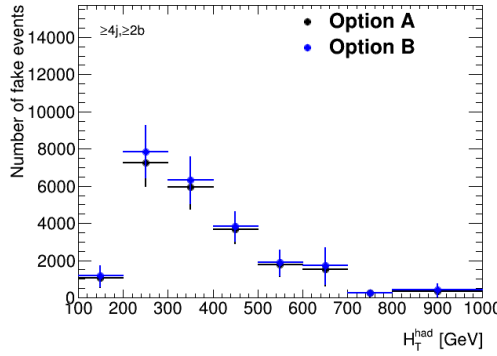


Figure 7.32: Distribution of the H_{T}^{had} in the validation region, showing the number of fake events. The fake estimate in *option A* (black) is parametrized as function of lepton p_T , $\min\Delta R(l, \text{jet})$, and lepton η , whereas the estimate in *option B* (red) is parametrized as function of leading jet p_T , $\min\Delta R(l, \text{jet})$, and lepton η .

2- The Choice of the Fake-enriched Region

The choice of the fake-enriched region to estimate the fake efficiencies was checked in the $e+\text{jets}$ channel. Figure 7.33 shows the fake efficiency as function of lepton p_T and the leading jet p_T . CR_1 is the nominal region used in the analysis and defined in Table 7.7, CR_2 is defined by requiring $m_T^W < 20$ GeV & $m_T^W + E_T^{\text{miss}} < 60$ GeV. The solid lines represent control regions with no requirement on the number of b -tagged jets, whereas the dashed lines represent control regions, which require events with at least one b -tagged jet.

A difference of about 20% is observed which corresponds to about 30% in the final fake estimate. This is considered as part of the systematic uncertainty on the final estimate of fake leptons.

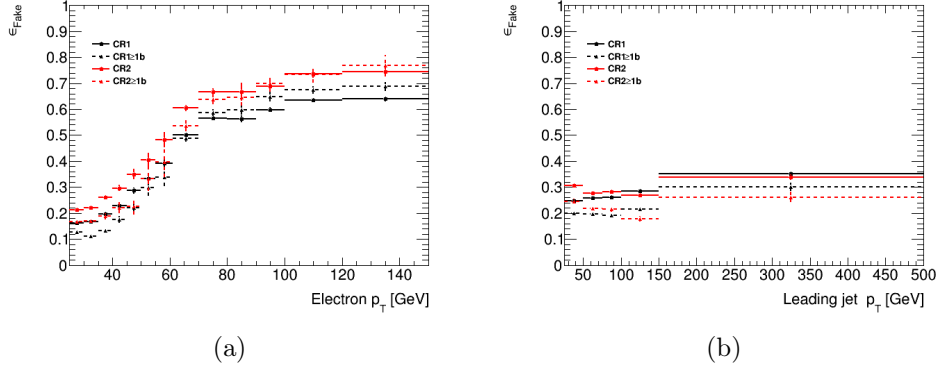


Figure 7.33: Fake ϵ_f efficiency as measured in the 2016 data in the e+jets channel, requiring at least two jets, as function of (a) the electron p_T and (b) leading jet p_T .

3- The Loose Sample

The low- p_T single-lepton triggers used in this analysis for the 2016 data taking, include isolation requirements on the candidate lepton, as shown in Table 7.1. This causes the Loose sample to be close to the Tight sample. Therefore, looser triggers were also studied. These triggers, listed in Table 7.11, have a looser lepton identification requirements, no isolation and no cut on the impact parameter (d_0). However, only a fraction (N) of the events satisfying the criteria is recorded using pre-scaled triggers, in order not to saturate the DAQ system. Since the trigger rate changes with instantaneous luminosity, dynamic pre-scales are used.

Fake efficiencies are expected to be different for events that are matched to the triggers, with or without isolation. Figure 7.34 illustrates the difference in the fake rate between the trigger used in the final results (in black) and the looser pre-scaled trigger (in red), as a function of lepton p_T . The looser pre-scaled triggers have about 20% lower fake rate. This difference is expected since the looser pre-scaled triggers (looser lepton identification requirement and no isolation requirement) trigger more fake leptons than the tighter triggers listed in Table 7.1.

Even though pre-scale triggers are looser, in regions with limited statistics, high

Event filter	Online object	p_T [GeV]	Pre-scale
e26_lhvloose_nod0_L1EM20VH	electron	26	from 20 up to 1000
mu_24	muon	24	50

Table 7.11: Loose pre-scaled low- p_T threshold single-lepton triggers used in the 2016 data taking. These triggers are not used in the final results. "Online" refers to the object used in the trigger logic. The electron identification operating point is represented by "lhloose" (see details in Section 5.2). "nod0" refers to absence of the track impact parameter requirement. "L1EM20VH" stands for the seed of lowest prescaled single electron trigger where "V" refers to the η -dependent threshold, "H" refers to the hadronic isolation.

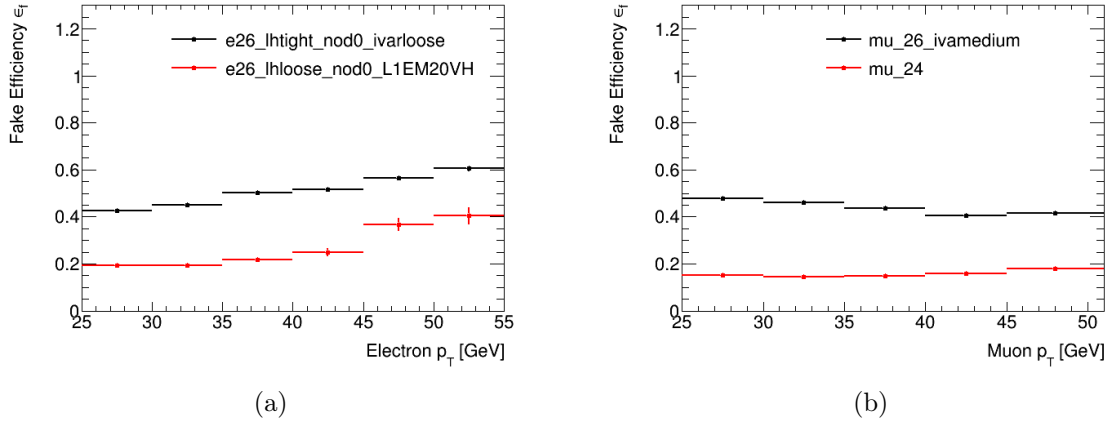


Figure 7.34: Fake efficiencies ϵ_f derived in the control regions in (a) e+jets and in (b) μ +jets for leptons fired by the nominal low- p_T triggers with isolation requirement (black) and the pre-scaled low- p_T triggers without isolation requirements (red) using the 2016 data.

prescales of order of 10^3 cause significant fluctuations in the event number, as shown in Figure 7.35 in blue. Events with significant fluctuations were found to have an expected weight obtained from the matrix method, while a prescale of about 100, resulting in an unrealistic increase in the fake estimate. To avoid these statistical fluctuations, the runs with very large prescales (above > 200) were removed. Keeping in mind that this is slightly over half of the data and then by scaling up the fake estimate accordingly, any event that passes these requirements and has a prescale of 100 would still cause statistical fluctuations, as shown in Figure 7.35 in black. Despite the fluctuations caused by the high pre-scale, the fake estimate in the most sensitive regions of the analysis is compatible with zero.

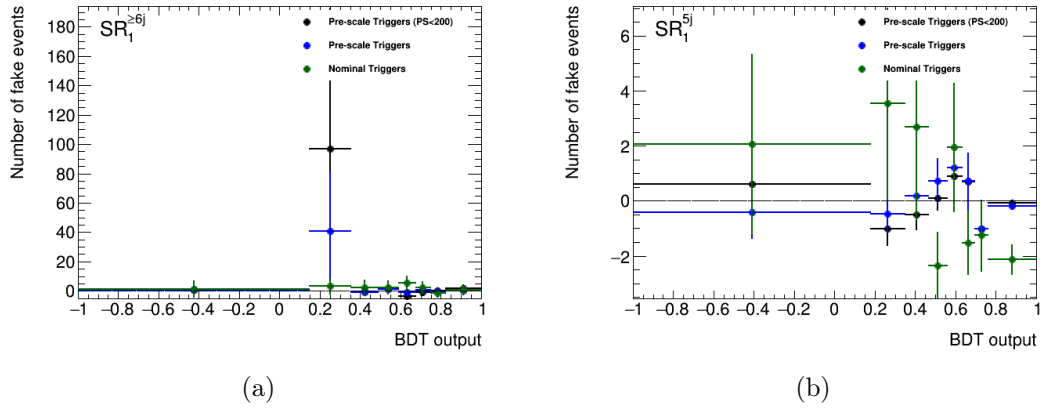


Figure 7.35: Distributions of the BDT output in (a) $SR_1^{>6J}$, and (b) SR_1^{5J} , showing the estimate of the fake background obtained from the matrix method. The error bars represent the statistical uncertainty. The distributions in black represent fake events selected using the pre-scaled trigger for low- p_T leptons but requiring the pre-scales to be < 200 and rescaling to the corresponding luminosity, in blue represent events selected using the pre-scaled trigger for low- p_T leptons, and in green events selected using the nominal triggers.

7.7.4 Other Backgrounds

The $t\bar{t}V$, single top (s-channel and Wt -channel), W/Z +jets, and diboson backgrounds are estimated from MC simulations as detailed in Section 7.3.3.

7.8 Kinematic Distributions in the Analysis Regions

The quality of the background modeling, explained in Section 7.7, has to be assessed by comparing the simulation with the measured data in control regions. Figure 7.36 shows the number of selected jets per event in the inclusive single-lepton selection. A very good modeling of the jet multiplicity is observed up to nine jets.

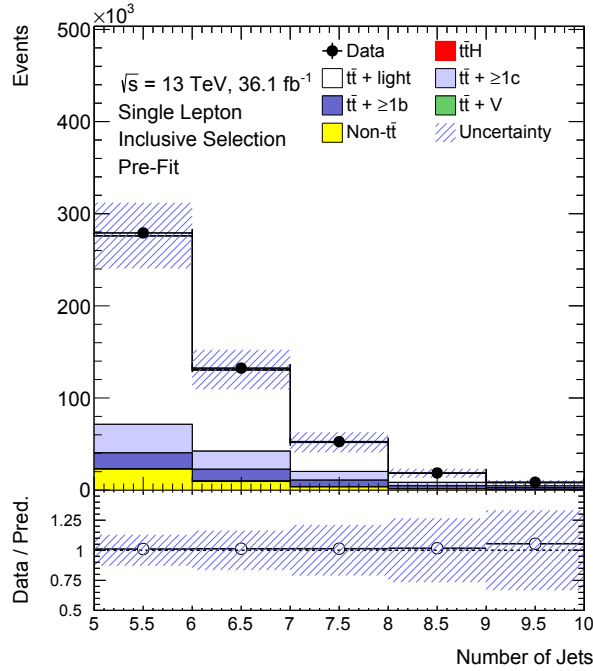


Figure 7.36: Comparison of the predicted number of jets to the one observed in data in the inclusive single-lepton channel selection. The hashed area represent the sum of the statistical and systematic uncertainties. Distributions are shown before the fit procedure, uncertainties on the normalisation of $t\bar{t} + \geq 1b$ or $t\bar{t} + \geq 1c$ are not included. Backgrounds from non- $t\bar{t}$ processes are grouped together and represented in yellow.

Figure 7.37 shows the number of b -tagged jets using the four operating points of the MV2c10 tagger. Clear slopes are observed and MC simulations tend to underestimate the data at high b -tagged multiplicities. A discrepancy of 10% up to 20% is seen. However, the

observed differences between data and MC simulations are within the assigned uncertainties on the prediction.

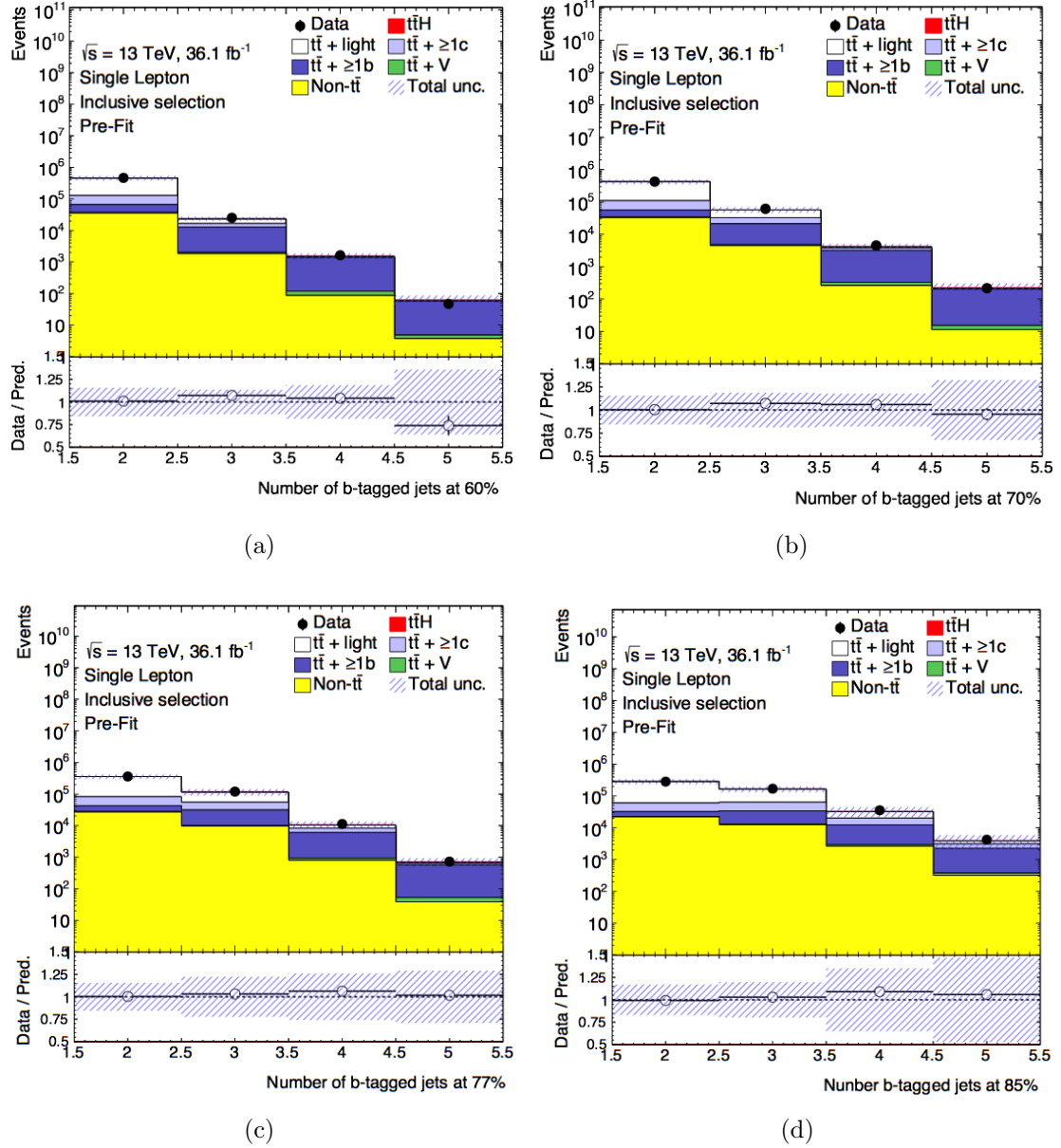


Figure 7.37: Comparison of the predicted number of b -tagged jets to the one observed in data for the four operating points of the MV2c10 tagger in the inclusive single-lepton channel selection. The hashed area represent the sum of the statistical and systematic uncertainties. Distributions are shown before the fit procedure, uncertainties on the normalisation of $t\bar{t} + \geq 1b$ or $t\bar{t} + \geq 1c$ are not included. Backgrounds from non- $t\bar{t}$ processes are grouped together and represented in yellow.

Figure 7.38 shows the electron and muon p_T , η , and ϕ in the five-jet $t\bar{t}$ +light control region in the single-lepton channel. Similar distributions for the leading jet are shown in Figure 7.39, along with aplanarity, H_{had}^T , and missing E_T . Overall, a good agreement is observed and data agrees with MC within the assigned systematic uncertainties. A small difference is present at low p_T , as seen in Figure 7.39 (a) and the first two bins in the H_{had}^T distributions in Figure 7.39 (e), where the data deviates from MC but still within the assigned systematic uncertainties.

Similar distributions are shown in Figure 7.40, and Figure 7.41, for the most sensitive regions ($SR_1^{\geq 6j}$) in the six-jet single-lepton channel.

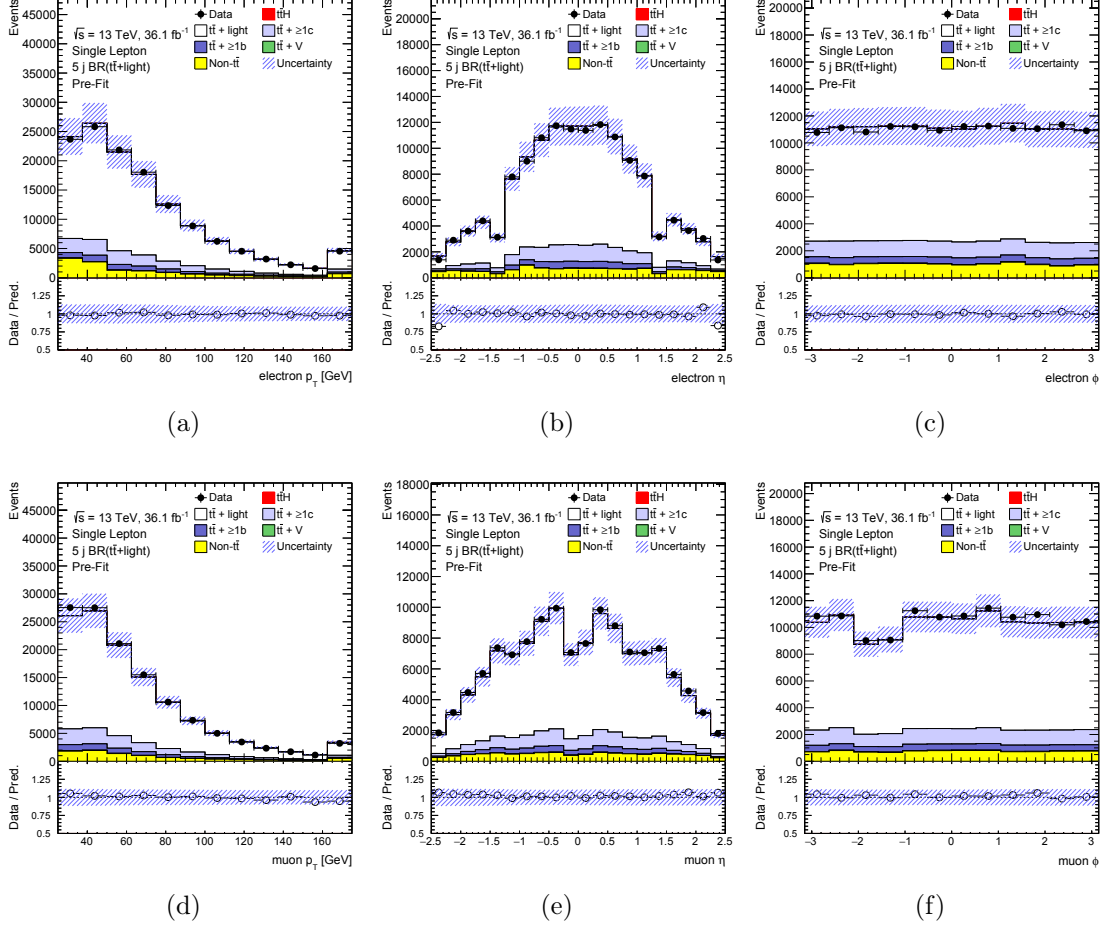


Figure 7.38: Comparison of the predicted and observed p_T , η , and ϕ of the (a-c) electron and (d-f) muon distributions of the five-jet $t\bar{t}$ +light control region in the single-lepton channel. The hashed area represent the sum of the statistical and systematic uncertainties. Distributions are shown before the fit procedure, uncertainties on the normalisation of $t\bar{t} + \geq 1b$ or $t\bar{t} + \geq 1c$ are not included. Backgrounds from non- $t\bar{t}$ processes are grouped together and represented in yellow.

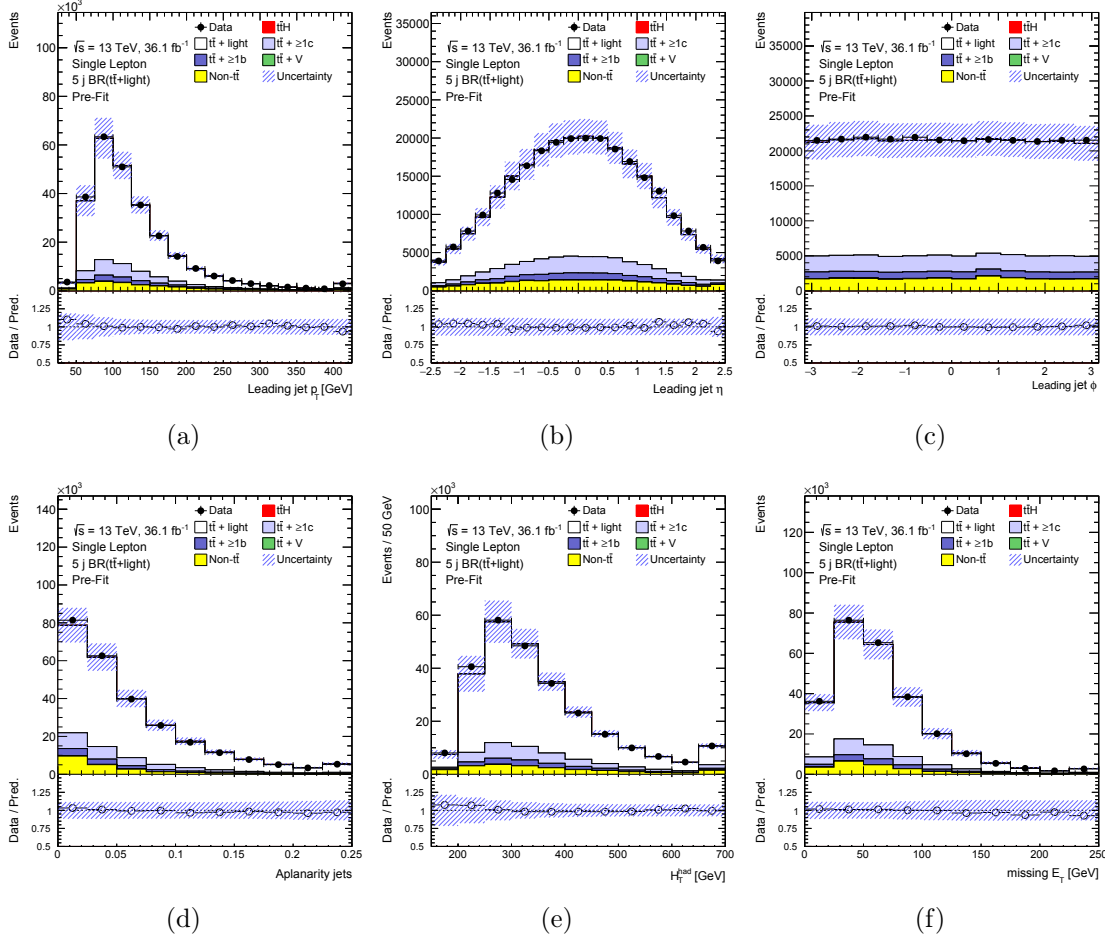


Figure 7.39: Comparison of the predicted and observed p_T , η , and ϕ of the (a-c) leading jet, (d) aplanarity, (e) H_{had}^T , and (f) missing E_T of the five-jet $t\bar{t}$ +light control region in the single-lepton channel. The hashed area represent the sum of the statistical and systematic uncertainties. Distributions are shown before the fit procedure, uncertainties on the normalisation of $t\bar{t}$ + $\geq 1b$ or $t\bar{t}$ + $\geq 1c$ are not included. Backgrounds from non- $t\bar{t}$ processes are grouped together and represented in yellow.

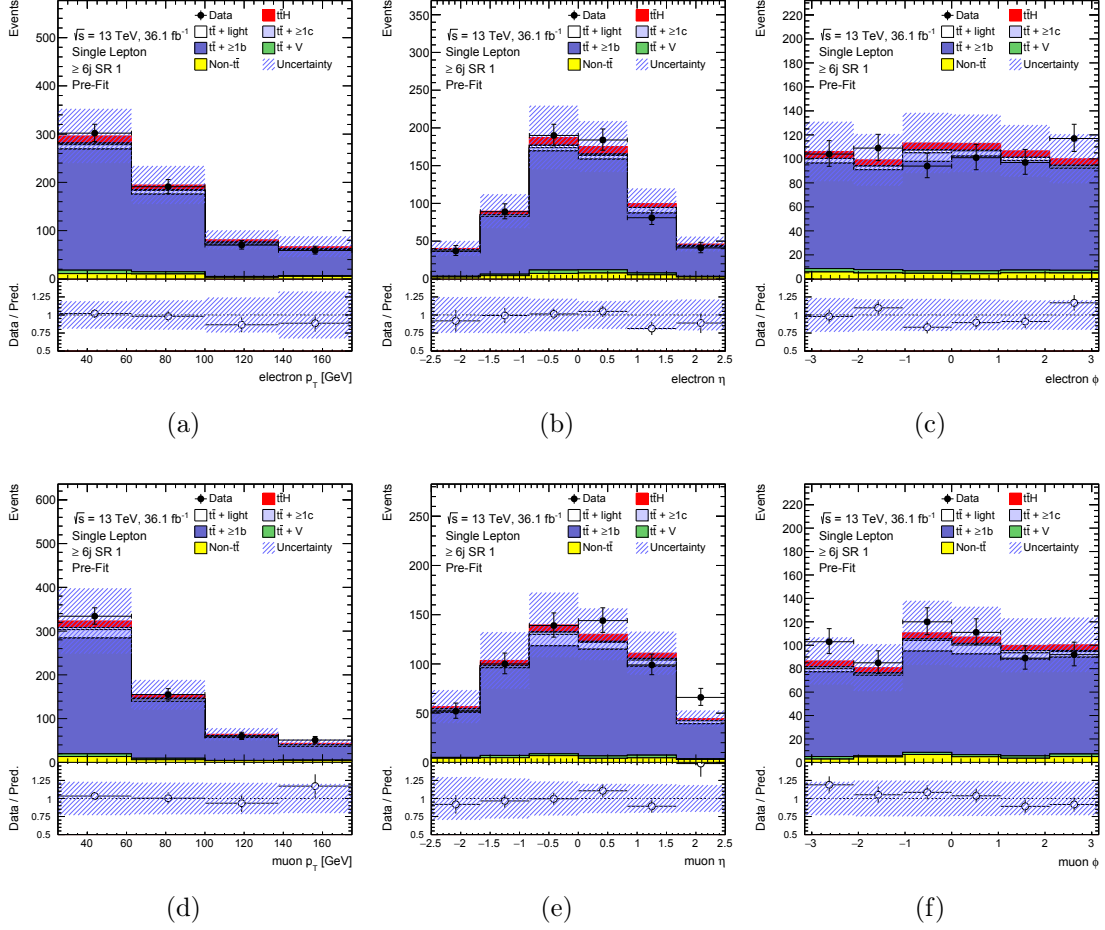


Figure 7.40: Comparison of the predicted and observed p_T , η , and ϕ of the (a-c) electron and (d-f) muon distributions of the six-jet signal region ($SR_1^{\geq 6j}$) in the single-lepton channel. The hashed area represent the sum of the statistical and systematic uncertainties. Distributions are shown before the fit procedure, uncertainties on the normalisation of $t\bar{t} + \geq 1b$ or $t\bar{t} + \geq 1c$ are not included. Backgrounds from non- $t\bar{t}$ processes are grouped together and represented in yellow.

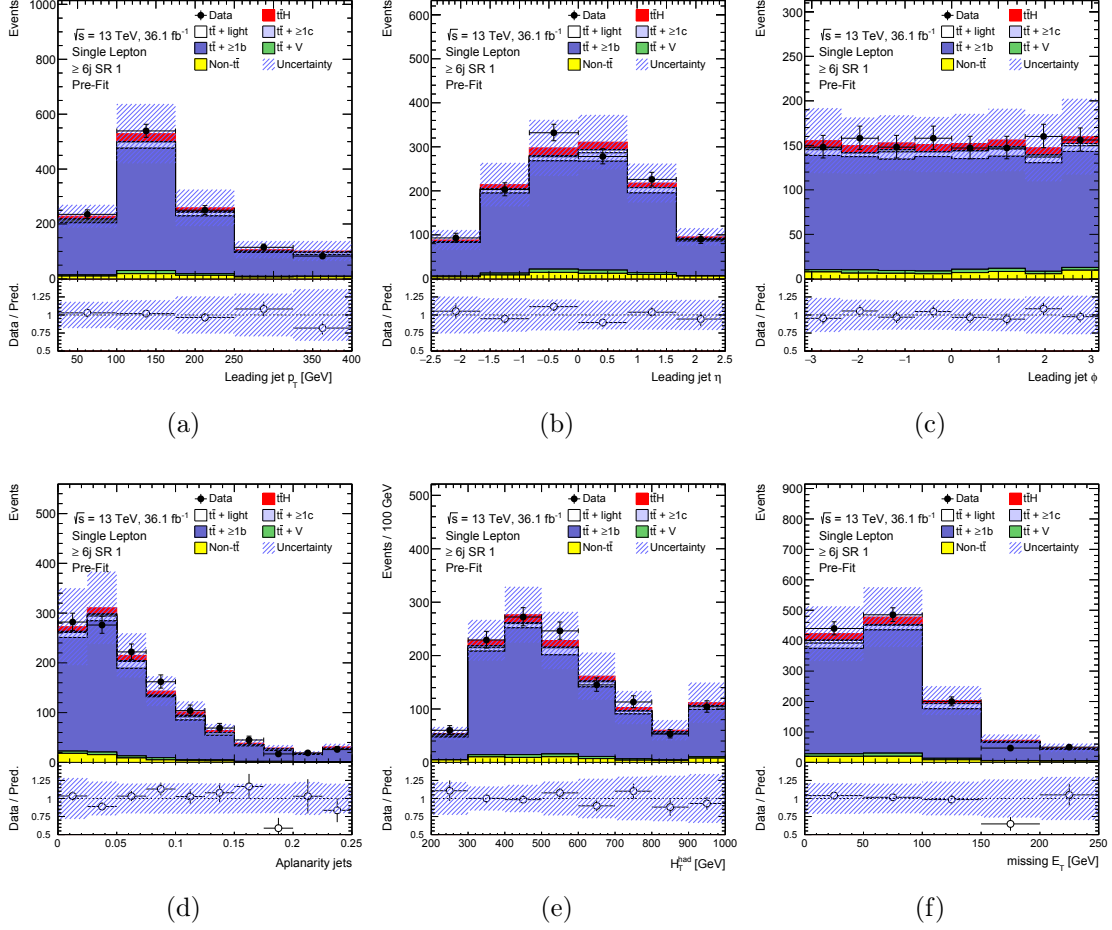


Figure 7.41: Comparison of the predicted and observed p_T , η , and ϕ of the (a-c) leading jet, (d) aplanarity, (e) H_{had}^T , and (f) missing E_T of the six-jet signal region ($SR_1^{\geq 6j}$) in the single-lepton channel. The hashed area represent the sum of the statistical and systematic uncertainties. Distributions are shown before the fit procedure, uncertainties on the normalisation of $t\bar{t} + \geq 1b$ or $t\bar{t} + \geq 1c$ are not included. Backgrounds from non- $t\bar{t}$ processes are grouped together and represented in yellow.

7.9 Systematic Uncertainties

Various sources of systematic uncertainty affect the search presented in this thesis, including those related to the luminosity, the identification and reconstruction of the physics objects, and the MC simulation of the signal and background processes. In the following, a brief description of the sources of systematic uncertainty will be provided together with their size. A particular emphasis will be made on the ones related to the $t\bar{t}$ background prediction, which will be seen to have the largest impact on the sensitivity of the measurement. The systematic variations can affect the amount of signal and background estimated in the different regions as well as the shape of the final discriminant distributions.

7.9.1 Experimental Uncertainties

The uncertainty on the integrated luminosity of the combined 2015 + 2016 dataset is 2.1%. It is derived following the methodology detailed in [155]. This uncertainty is applied to the normalization of all processes determined by MC simulations.

An uncertainty is considered on the re-weighting of the pileup distributions. This pileup re-weighting is applied in order to correct for the differences in the pileup distributions between MC simulation and data.

The $t\bar{t}H$ measurement is based on the reconstructed objects, leptons and jet. The identification efficiencies are derived in simulation and are corrected with scale factors to match the data. Therefore, the uncertainties on these corrections have to be considered. The correction related to scale factors applied on efficiencies for triggering, reconstructing, and identifying objects, is applied by modifying the event weight. Whereas the correction related to the energy scales and resolutions, is applied by smearing or re-scaling the energies of the objects.

Lepton-related uncertainties correspond to the electron and muon reconstruction, identification, trigger, isolation efficiencies, and the resolution of the measurement of the energy and momentum. A total of 24 independent components are considered. These uncertainties are below 1% and have a negligible effect on the analysis.

Jet-related uncertainties are related to the jet energy resolution (JER), jet energy scale (JES), and jet vertex tagger (JVT). A total of twenty independent sources are considered for the JES uncertainties. These uncertainties arise from in-situ calibration

techniques and corrections derived from MC which include statistical, detector, modeling effects, jet flavor, pileup corrections, η dependence, and high- p_T jets calibration. Although the uncertainties on JES for an individual jet are not large, about 5.5% for jets with $p_T = 25$ GeV and below 1.5% for central jets with p_T in the range of $\simeq 100$ GeV – 1.5 TeV, the effects are amplified by the large number of jets in the final state.

b -tagging uncertainties include the uncertainties from the b -tagging efficiency of jets as well as the mis-tagging efficiency of c - and light-jets. These uncertainties are a mixture of statistical, experimental and modeling uncertainties that are split into orthogonal sub-components. The efficiency to correctly tag b -jets is measured in data using dilepton $t\bar{t}$ events. The mis-tag rate for c -jets is determined from c -jets of the hadronic W decay in $t\bar{t}$ events [137], and the one for light jets is measured in multi-jet events using jets that contain secondary vertices and tracks with impact parameters consistent with a negative lifetime [193]. First, the b -tagging efficiencies and the mis-tag rates are derived for the four b -tagging operating points used in the analysis as a function of the jet kinematics. Then, they are combined into a distribution with the corresponding uncertainties that correctly describe correlations across multiple working points. The uncertainty associated with the b -tagging efficiency has 30 independent sources, while 20 (80) independent sources are associated with c -jet (light-jets) mis-tag rates.

Missing transverse energy-related uncertainties are affected by the uncertainties associated with leptons and jet energy scales and resolution. Since the analysis does not make direct use of the E_T^{miss} information in the selection but only in the event reconstruction, its uncertainties have a typically small effect below 0.5% on the analysis. Additional Uncertainties in the scale and resolution of the soft term are considered, for a total of three additional sources of systematic uncertainty.

7.9.2 *Uncertainties Related to the Background Estimation*

The uncertainties on the cross-section for simulated samples, which are listed in Table 7.3, are taken from the latest available theoretical calculations and only affect the normalization. Comparisons between different MC samples probe various aspects of the event modeling and can be used to assess the modeling uncertainties associated with the search. They affect the normalization and/or the shape. The following details the systematic uncertainties

related to the background estimation.

Since the $t\bar{t}$ +jets represent by far the largest source of background, a large spectrum of uncertainties were considered. These include uncertainties associated with the choice of a particular MC prediction for the top-quark pair production of the matrix element, the extra radiation, the choice of parton shower and hadronization model, and uncertainties affecting the modeling of $t\bar{t}+ \geq 1b$ and $t\bar{t}+ \geq 1c$ production. These uncertainties are estimated from comparisons between the nominal $t\bar{t}$ sample and the alternative $t\bar{t}$ samples, which are listed in Table 7.2.

The different processes $t\bar{t}+ \geq 1b$, $t\bar{t}+ \geq 1c$, and $t\bar{t}$ +light are effected by different types of uncertainties. For example, the $t\bar{t}$ +light has additional diagrams and profits from relatively precise measurements in data, the $t\bar{t}+ \geq 1b$ and $t\bar{t}+ \geq 1c$ can have similar or different diagrams depending on the flavor scheme used in the PDF, and the differences in the mass of the c - and b -quarks. As a consequence, all uncertainties in the $t\bar{t}$ +jets background modeling are assigned independent parameters for the $t\bar{t}+ \geq 1b$, $t\bar{t}+ \geq 1c$, and $t\bar{t}$ +light processes.

The following uncertainties are designed to target one modeling component at a time, in order to minimize correlations among the different MC models. The uncertainties associated with the choice of NLO generator are estimated by comparing the predictions of POWHEG-BOX+PYTHIA 8 to SHERPA5F. The choice of parton shower and hadronization model are estimated from comparing the prediction of PYTHIA 8 to HERWIG 7. The final state radiation (ISR/FSR) is assessed with two alternative POWHEG-BOX+PYTHIA 8 samples produced with different amount of radiations, renormalization and factorization scales, h_{damp} parameter, and the Var3c variation of the A14 parameter set. The settings used to generate these samples are detailed in Section 7.3.3.

For the $t\bar{t}+ \geq 1b$ process, the differences between the predictions from POWHEG-BOX+PYTHIA 8 and SHERPA4F are considered as one additional source of uncertainty. This uncertainty covers the difference between the description of the $t\bar{t}+ \geq 1b$ process by the NLO $t\bar{t}$ inclusive MC sample and a description at NLO of the $t\bar{t} + b\bar{b}$ process in the matrix element. However, this uncertainty is not applied to the $t\bar{t} + b(\text{MPI/FRS})$ sub-category since it is not included in the SHERPA4F sample. A normalization uncertainty of 50% is applied for the contribution from MPI based on studies of different underlying events sets of tuned parameters.

The above described uncertainties do not affect the relative fractions of $t\bar{t} + \geq 1b$ sub-categories ($t\bar{t} + b$, $t\bar{t} + B$, $t\bar{t} + b\bar{b}$, and $t\bar{t} + \geq 3b$), which are re-weighted to match the prediction of SHERPA4F. The uncertainties in these fractions in SHERPA4F are assessed separately for each independent source. The POWHEG-BOX+PYTHIA 8 is re-weighted to the SHERPA4F and its variations and the differences are taken as part of this uncertainty. The uncertainties from these SHERPA variations are shown in the red band in Figure 7.15. The large difference between the POWHEG-BOX+PYTHIA 8 and the SHERPA4F in the $t\bar{t} + \geq 3b$ process, is not covered by the considered uncertainties. Therefore, this sub-process is given an extra 50% normalization uncertainty.

For the background arising from $t\bar{t} + \geq 1c$, there is less guidance from theory or experiment to determine the best approach in handling the production of charm jets; in the parton shower compared to the prediction with $t\bar{t} + c\bar{c}$ calculated at NLO in the matrix element. In order to estimate the uncertainty on the production of charm jets, a dedicated $t\bar{t} + c\bar{c}$ sample is generated with NLO prediction in the matrix element, including massive c -quarks (using the 3F scheme for the PDFs). This sample is produced with MG5_aMC@NLO interfaced with HERWIG++, as described in [194]. This additional uncertainty on the $t\bar{t} + \geq 1c$ prediction arises from the difference between the mentioned sample and the inclusive $t\bar{t}$ sample produced with the same generator and a 5F scheme PDF set, in which the $t\bar{t} + \geq 1c$ process only originated through the parton shower.

Table 7.12 summarises the systematic uncertainties affecting the prediction of the $t\bar{t}$ +jets background.

Uncertainties on the data-driven fake lepton background arise from the limited sample size in data, particularly in the analysis phase space at high jet and b -tag multiplicity, as well as a systematic uncertainty related to the lepton misidentification rate measurements in various fake-enriched control regions and the choice of parametrization of both the real and fake efficiencies. A combined normalization uncertainty of 50% is assigned to the overall estimated yield of fake lepton events in the single-lepton channel. This uncertainty is considered to be uncorrelated between the electron and muon channel, and between the analysis regions with exactly five jets and those with at least six jets. In the dilepton channel, a 25% uncertainty is assigned for the fake lepton background, correlated across lepton flavors and all analysis regions.

Non- $t\bar{t}$ simulated background processes such as W/Z +jets, diboson, and single top,

Systematic source	Description	$t\bar{t}$ categories
$t\bar{t}$ cross-section	From NNLO+NNLL calculation	All, correlated
$k(t\bar{t}+ \geq 1c)$	$t\bar{t}+ \geq 1c$ normalization	$t\bar{t}+ \geq 1c$
$k(t\bar{t}+ \geq 1b)$	$t\bar{t}+ \geq 1b$ normalization	$t\bar{t}+ \geq 1b$
SHERPA5F vs. nominal	Related to the choice of the NLO generator	All, uncorrelated
PS & hadronization	POWHEG-BOX+HERWIG 7 vs. POWHEG-BOX+PYTHIA 8	All, uncorrelated
ISR / FSR	Variations of μ_R , μ_F , h_{damp} and A14 Var3c parameters	All, uncorrelated
$t\bar{t}+ \geq 1c$ ME vs. inclusive	MG5_AMC@NLO+HERWIG++: ME prediction (3F) vs. incl. (5F)	$t\bar{t}+ \geq 1c$
$t\bar{t}+ \geq 1b$ SHERPA4F vs. nominal	Comparison of $t\bar{t}+b\bar{b}$ NLO (4F) vs. POWHEG-BOX+PYTHIA 8 (5F)	$t\bar{t}+ \geq 1b$
$t\bar{t}+ \geq 1b$ renorm. scale	Up or down by a factor of two	$t\bar{t}+ \geq 1b$
$t\bar{t}+ \geq 1b$ resumm. scale	Vary μ_Q from $H_T/2$ to μ_{CMMPs}	$t\bar{t}+ \geq 1b$
$t\bar{t}+ \geq 1b$ global scales	Set μ_Q , μ_R , and μ_F to μ_{CMMPs}	$t\bar{t}+ \geq 1b$
$t\bar{t}+ \geq 1b$ shower recoil scheme	Alternative model scheme	$t\bar{t}+ \geq 1b$
$t\bar{t}+ \geq 1b$ PDF (MSTW)	MSTW vs. CT10	$t\bar{t}+ \geq 1b$
$t\bar{t}+ \geq 1b$ PDF (NNPDF)	NNPDF vs. CT10	$t\bar{t}+ \geq 1b$
$t\bar{t}+ \geq 1b$ UE	Alternative set of tunable parameters for the underlying event	$t\bar{t}+ \geq 1b$
$t\bar{t}+ \geq 1b$ MPI	Up or down by 50%	$t\bar{t}+ \geq 1b$
$t\bar{t}+ \geq 1b$ normalization	Up or down by 50%	$t\bar{t}+ \geq 1b$

Table 7.12: Summary of the sources of systematic uncertainties on the $t\bar{t}$ +jets modelling. The systematic uncertainties listed in the second section of the table are evaluated independently of the categorization of the events into $t\bar{t}+ \geq 1b$ and $t\bar{t}+ \geq 1c$. The third section of the table lists the systematic uncertainties that affect the $t\bar{t}+ \geq 1b$ sub-categories. The last column of the table states the $t\bar{t}$ category to which a systematic uncertainty is applied. In the cases where all the three categories of $t\bar{t}$ +light, $t\bar{t}+ \geq 1c$, $t\bar{t}+ \geq 1b$ are involved, the word (All) is listed. Also, the last column indicates whether the uncertainty is considered as correlated or uncorrelated across $t\bar{t}$ categories.

represent a minor fraction of the total background; in the control region $CR_{t\bar{t}+light}^{\geq 6j}$ they account for $\approx 7\%$ of the background, while with increasing signal purity, their contribution becomes smaller. In the most sensitive signal region ($SR_1^{\geq 6j}$), their yields are comparable to the ones expected for the SM Higgs boson signal. Therefore, a less refined treatment of the uncertainties associated with these small backgrounds was adopted after verifying that the implemented uncertainties have a sub-leading effect on the sensitivity of the analysis.

A conservative uncertainty of 40% is assumed for the W +jets cross-section, with an additional 30% normalization uncertainty used for W +heavy flavour jets component, taken as uncorrelated between events with at least two heavy-flavor jets. These uncertainties are based on variations of the factorization and renormalization scales and of the matching parameters in the SHERPA simulation.

An uncertainty of 35% is applied to the Z +jets normalization, uncorrelated across jet bins. This accounts for both the variation of the scales and matching parameters in the SHERPA simulation. A correction factor of 1.3 is applied to the Z boson production associated to at least one heavy-flavour jet.

An uncertainty of $^{+5\%}_{-4\%}$ is considered for each of the three single-top production mode cross-sections [179–181]. For the Wt and t -channel production modes, uncertainties on the choice of the parton shower, hadronization model, initial and final-state radiation is evaluated according to a set of alternative samples in a manner similar to that used for the $t\bar{t}$ process. The nominal prediction is compared to samples generated with POWHEG-BOX interfaced with HERWIG++ and with alternative POWHEG-BOX+PYTHIA 6 samples with factorization and renormalization scale variations and appropriate variations of the Perugia 2012 set of tunable parameters. Additional uncertainties on the interference between Wt and $t\bar{t}$ production in the NLO [178] calculation is assessed by comparing the default "diagram removal" scheme to an alternative "diagram subtraction" scheme.

A 50% normalisation uncertainty on the diboson background is assumed, which includes uncertainties on the inclusive cross-section and an additional jet production [195].

The theoretical uncertainty on the $t\bar{t}V$ NLO cross-section is 15% [196]. An uncertainty associated with the choice of the generator is derived by comparing the nominal sample to the alternative samples generated using SHERPA.

7.9.3 *Uncertainties on the Signal Modeling*

The systematic uncertainties related to the modeling of the $t\bar{t}H$ process are assessed by varying the parameter settings in the simulation of the parton shower and hadronization. A theoretical uncertainty of $^{+5.8\%}_{-9.2\%}(\text{scale}) \pm 3.6\%(\text{PDF})$ is applied on the cross-section of the $t\bar{t}H$ signal, the first component representing the QCD scale uncertainty and the second the $\text{PDF}+\alpha_S$ uncertainty [18, 160–164]. Uncertainty on the shape of the distribution due to the QCD scale choice is estimated by varying the renormalization and factorization scales. The uncertainty related to the showering and hadronization is estimated by comparing the nominal prediction from MG5_aMC@NLO+PYTHIA 8 to the one from MG5_aMC@NLO interfaced to HERWIG++. Lastly, uncertainties on the Higgs boson branching ratios are also considered; these amount to 2.2% for the $b\bar{b}$ decay mode [18].

7.9.4 *Summary of Systematic Uncertainties*

Table 7.13 lists all the considered systematic uncertainties, indicates if they affect only the normalization or both the shape and normalization, and lists the number of components

to parametrize the uncertainty.

Systematic uncertainty	Type	Components
Luminosity	N	1
Reconstructed Objects		
Electron trigger+reco+ID+isolation	SN	4
Electron energy scale+resolution	SN	2
Muon trigger+reco+ID+isolation	SN	10
Muon momentum scale+resolution	SN	5
Taus detector, insitu and model	SN	3
Pileup modelling	SN	1
Jet vertex tagger	SN	1
Jet energy scale	SN	20
Jet energy resolution	SN	2
Missing transverse energy scale+resolution	SN	3
b -tagging efficiency	SN	30
c -mistag rate	SN	20
Light-mistag rate	SN	60
Mistag extrapolation $c \rightarrow \tau$	SN	1
Background and Signal Model		
$t\bar{t}$ cross section	N	1
$t\bar{t} + \geq 1c$: normalisation	N	1
$t\bar{t} + \leq 2b$: normalisation	N	1
$t\bar{t} + \geq 3b$: normalisation	N	1
$t\bar{t} + \geq 1b$: NLO Shape	SN	9
$t\bar{t} + \geq 1c$: NLO Shape	SN	1
$t\bar{t} + \geq 1b$: 4F vs 5F Shape	S	1
$t\bar{t}$ modelling: residual Radiation	SN	3
$t\bar{t}$ modelling: residual NLO generator	SN	3
$t\bar{t}$ modelling: residual parton shower+hadronisation	SN	3
W +jets normalisation	N	3
Z +jets normalisation	N	3
Single top cross section	N	1
Single top model	SN	2
Diboson normalisation	N	1
Fakes normalization	SN	6
$t\bar{t}V$ cross section	N	4
$t\bar{t}V$ modelling	SN	2
tZ cross section	N	2
tWZ cross section	N	1
$t\bar{t}WW$ cross section	N	2
4-tops cross section	N	1
$tHjb$ cross section	N	3
WtH cross section	N	2
$t\bar{t}H$ cross section	N	2
$t\bar{t}H$ branching ratios	N	3
$t\bar{t}H$ modelling	SN	1

Table 7.13: The list of systematic uncertainties considered. "N" means that the uncertainty is taken as normalization-only for all processes and channels affected, whereas "SN" means that the uncertainty is taken on both shape and normalization. Some of the systematic uncertainties are split into several components for a more accurate treatment.

7.10 Statistical Analysis and Results

All analysis regions are combined in a statistical model using a profile likelihood fit in order to test for the presence of a $t\bar{t}H$ signal, assuming a Higgs boson mass of $m_H = 125$ GeV. The profile likelihood function is described in Section 7.10.1, where the impact of the systematic uncertainties is propagated through the so-called *nuisance parameters* (NP) in the fit. The fit model is described in Section 7.10.2. The performance of the $t\bar{t}H(H \rightarrow b\bar{b})$ analysis and the validation of the fit model are evaluated from a fit to the Asimov data-set [197], as described in Section 7.10.3. This section, also summarizes the fit to pseudo-data built from an alternative $t\bar{t}$ model in order to evaluate the quality and the completeness of the systematic model. Section 7.10.4 presents the results obtained from the fit to data. Finally, Section 7.10.5 explains the procedure leading to the determination of the signal significance and the upper limit of the signal strength of $t\bar{t}H(H \rightarrow b\bar{b})$ production.

7.10.1 The Profile Likelihood Fit

The distributions of the discriminants from each of the analysis regions are combined in a profile likelihood fit to test for the presence of a $t\bar{t}H$ signal and to constrain the backgrounds. For each bin i of the input distribution of each region r , the number of data events $N_{r,i}^{\text{data}}$ are compared to the expected bin content $N_{r,i}^{\text{exp}}$. The expected bin content $N_{r,i}^{\text{exp}}$ is expressed as the following:

$$N_{r,i}^{\text{exp}}(\mu, k_1, \dots, k_m, \theta_1, \dots, \theta_n) = \mu \cdot N_{r,i,\text{sig}}^{\text{exp}}(\theta_1, \dots, \theta_{n_{\text{sig}}}) + \sum_{b \in \text{bkg}} k_b \cdot N_{r,i,b}^{\text{exp}}(\theta_1, \dots, \theta_{n_b}), \quad (7.9)$$

where n is the total number of nuisance parameters, $(\theta_1, \dots, \theta_{n_i})$ is the set of n_i nuisance parameters related to the sample i being signal (sig) or background (bkg), k_b is the normalization factor on the background b and is referred to as a k -factor, m is the number of backgrounds, and $\mu = \sigma_{t\bar{t}H}/\sigma_{t\bar{t}H}^{\text{SM}}$ is the signal strength. In the following, k is used for the set of all normalization factors and θ for the set of all nuisance parameters. For every systematic variation listed in Table 7.13, there is a nuisance parameter θ_i that modifies the shape and/or the normalization of the templates depending on the parametrized systematic uncertainty. The templates are formed from the expected distributions where the nuisance

parameters are varied. The normalization factors (k -factors) and the signal strength (μ), modify only the normalization of the template distributions.

The data content of each bin is expected to follow a Poisson probability. Therefore, the primary likelihood function $\mathcal{L}_{\text{main}}(\mu, k, \theta)$ is constructed as the product of a Poisson probability terms for each bin:

$$\mathcal{L}_{\text{main}}(\mu, k, \theta) = \prod_{r \in \text{regions}} \prod_{i \in \text{bins}} \frac{(N_{r,i}^{\text{exp}}(\mu, k, \theta))^{N_{r,i}^{\text{data}}}}{N_{r,i}^{\text{data}}!} \cdot e^{-N_{r,i}^{\text{exp}}(\mu, k, \theta)}. \quad (7.10)$$

The value $\theta = 0$, by construction, corresponds to the best knowledge of a specific parameter (nominal value). Uncertainty variations up to $\pm 1\sigma$ correspond to the 1σ uncertainty. The nuisance parameters are defined by the extrapolation ($|\theta| > 1$) and interpolation ($|\theta| < 1$) functions with constraints that $\theta = 0$ corresponds to no corrections and $\theta = \pm 1$ shifts the distribution by $\pm 1\sigma$ systematic uncertainty. A linear and exponential extrapolations for the shape and normalization components of the systematic uncertainties are used, respectively [198]. Therefore, two polynomial functions are defined, one for the shape and one for the normalization components of the systematic uncertainties. The deviation of the nuisance parameters, the normalization factors, or μ are referred to as a *pull*. Nuisance parameters are implemented using Gaussian constraints reflecting the prior knowledge of the systematic uncertainty and the likelihood function can be expressed as the following:

$$\mathcal{L}(\mu, k, \theta) = \mathcal{L}_{\text{main}}(\mu, k, \theta) \cdot \prod_{t=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta_t^2}{2}}. \quad (7.11)$$

The best estimate for the parameter set (μ, k, θ) is obtained by maximizing the likelihood function or by minimizing the negative log likelihood $-\log L$. The results presented here are obtained using the minimization as implemented in the minuit2 package of the ROOFIT framework [199–201]. The effects before performing the fit are referred to as "pre-fit", and after performing the fit are referred to as "post-fit".

7.10.2 The Fit Model

The fit model is described by the chosen variables to build the template distributions, and the list of systematic uncertainties and their correlations across the defined analysis regions. The distributions of the multivariate discriminant, *classification BDT output* as shown in Figure 7.13, from each of the signal regions of the analysis are combined in the profile likelihood fit to test for the presence of a $t\bar{t}H$ signal, while simultaneously determining the normalization and constraining the differential distributions of the dominant background components. In the control regions, only the event yield is used, with the exception of $CR_{t\bar{t}+\geq 1c}^{5j}$ and $CR_{t\bar{t}+\geq 1c}^{\geq 6j}$, where the $H_{T^{\text{had}}}$ distribution is used. The fit is performed on the combined events from the single-lepton channel and the dilepton channel and is performed simultaneously in all the nineteen analysis regions. Figure 7.42 shows the predicted number of events compared to the amount of observed data events, in each analysis region in the single-lepton and dilepton channels. Data overshoot the prediction in various regions with large fractions of $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ backgrounds. However, the difference lies within the systematic uncertainty band.

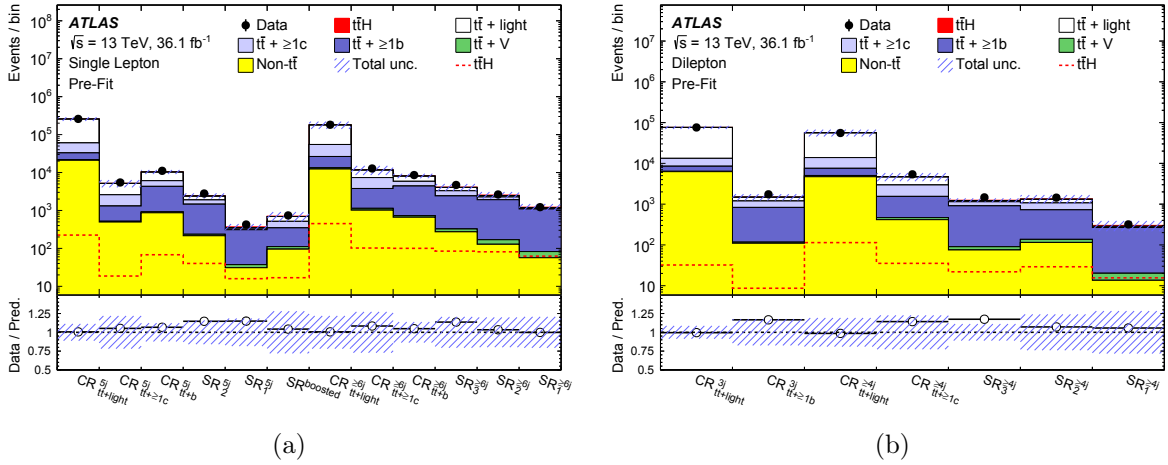


Figure 7.42: Comparison of predicted and observed event yields in all 19 analysis region, in the (a) single-lepton and (b) dilepton channels. These plots are shown before performing the fit to data. The hashed band represent the sum of the statistical and systematic uncertainties. Uncertainties on the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ background normalizations are not included as they are free-floating parameters in the fit.

The described fit uses the background dominated regions in order to improve the knowledge of the background, through constraints of the nuisance parameters or the

resulting correlations.

In control regions where the shape of the H_T^{had} distribution is not well modeled, such as the five- and six-jets $t\bar{t}$ +light enriched and $t\bar{t} + b$ enriched regions, only one bin is used in the fit to data. Figure 7.43 shows the H_T^{had} distribution in the five- and six-jets $t\bar{t}$ +light enriched regions in the single lepton channel. This is done to avoid propagating mismodeled effects from the control regions to the signal regions, and to avoid arbitrary pulls of the nuisance parameters in order to correct such mismodeling.

The signal strength is considered as a freely floating parameter, applied without any prior in the fit, along with the normalization factors of the $t\bar{t} + \geq 1b$ and $t\bar{t} + \geq 1c$ backgrounds. Studies have shown that MC simulations underestimate the $t\bar{t}$ +HF fractions with respect to data. Therefore, the normalization of the $t\bar{t}$ +HF components are free-floating in the fit and are applied without any prior. The k -factors defined in Equations 7.9-7.11 refer to the $t\bar{t} + \geq 1b$ and $t\bar{t} + \geq 1c$ backgrounds. None of the other backgrounds has a free-floating normalization; their normalization is controlled through specific nuisance parameters that reflect the theoretical knowledge of the respective cross sections or the uncertainty on the estimate methods. Such is the case for the fake lepton background estimated using the matrix method.

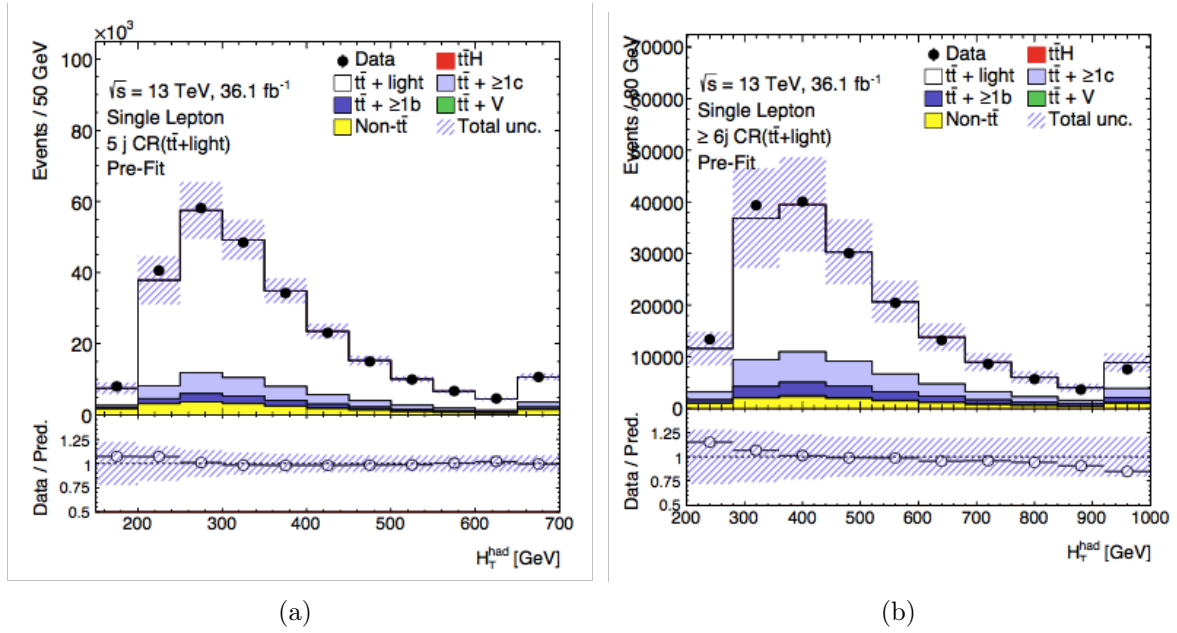


Figure 7.43: Comparison between data and prediction for the H_T^{had} distributions in the five- and six-jet $\text{tt}+\text{light}$ -enriched control regions in the single lepton channel. The distributions are the input to the fit to data. The hashed band represent the sum of the statistical and systematic uncertainties. Uncertainties on the $\text{tt} + \geq 1b$ and $\text{tt} + \geq 1c$ background normalizations are not included as they are free floating parameters in the fit.

7.10.3 Expected Performance

The fit model is validated and the performance of the $t\bar{t}H(H \rightarrow b\bar{b})$ analysis in terms of expected sensitivity and background constraints are evaluated from a fit to the Asimov data-set [197]. The Asimov data-set is built from the predicted distribution, in which a Poisson error in each bin corresponding to the statistical uncertainty of the data is assumed. Many studies and decisions in defining the final analysis strategy were based on the obtained performance in the Asimov fits, such as understanding constraints on a set of nuisance parameters and their impact on the signal sensitivity, the choice of the number of bins in the used distributions, and the choice of the BDTs. A selection of these studies are listed here for the single lepton channel.

Asimov Fits

The Asimov fit in the single lepton channel yields a signal strength of

$$\mu_{\text{Asimov data}} = 1.00 \pm 0.32(\text{stat})^{+0.60}_{-0.57}(\text{syst}) = 1.00^{+0.68}_{-0.65}, \quad (7.12)$$

which corresponds to a 1.5σ expected significance of the $t\bar{t}H(H \rightarrow b\bar{b})$ signal assuming a SM $t\bar{t}H$ production. The uncertainty on the signal strength is dominated by the systematic uncertainties and the data statistic only contributes to ± 0.32 of the uncertainty. Note that the total statistical uncertainty also includes the uncertainty on the normalizations of the $t\bar{t} + \geq 1b$ and $t\bar{t} + \geq 1c$ factors.

Table 7.14 shows the summary of the uncertainty impact on the signal strength error. The sources of uncertainty have been grouped into categories. The total statistical uncertainty is evaluated, by fixing all the nuisance parameters in the fit but the free-floating normalization factors for the $t\bar{t} + \geq 1b$, and $t\bar{t} + \geq 1c$ background components. The other quoted numbers are obtained by repeating the fit after having fixed a certain set of nuisance parameters corresponding to a group systematic uncertainty sources, and subtracting in quadrature the resulting total uncertainty on μ from the uncertainty from the full fit.

Despite the efforts to get the best possible predictions for the $t\bar{t} + b\bar{b}$ background by re-weighting the event fraction to match the fractions predicted by the 4F calculation, the sensitivity of the $t\bar{t}H(H \rightarrow b\bar{b})$ search is limited by the modeling of the dominant

$t\bar{t}+ \geq 1b$ background. The uncertainty on the signal strength associated with the $t\bar{t}+ \geq 1b$ background model nuisance parameters is $^{+0.49}_{-0.48}$ and in addition to an uncertainty of $^{+0.12}_{-0.14}$ arising from the $t\bar{t}+ \geq 1b$ normalization. Table 7.12 details the different sources of systematic that contribute to this large impact on the signal strength.

The second largest impact on the signal strength arises from the background MC statistics and from the statistical error on the fake lepton estimate. The combined impact on the signal strength is $^{+0.29}_{-0.31}$. Increasing the amount of MC generated events in the small phase space where the $t\bar{t}H$ signal is present, would significantly reduce the impact of this uncertainty on the signal strength.

The third largest impact on the signal strength arises from $t\bar{t}H$ modeling $^{+0.24}_{-0.03}$, which is highly asymmetric. Note that the applied theoretical uncertainty on the cross-section of the $t\bar{t}H$ signal is asymmetric, as mentioned in Section 7.9.4. Moreover, the seen asymmetry is intrinsic to the way this error is computed.

Uncertainty source	$\Delta\mu$	
$t\bar{t}+ \geq 1b$ modeling	+0.49	-0.48
Background model statistics	+0.29	-0.31
$t\bar{t}H$ modeling	+0.24	-0.03
Jet flavor tagging	+0.16	-0.15
Jet energy scale and resolution	+0.12	-0.13
$t\bar{t}+ \geq 1c$ modeling	+0.11	-0.12
Other background modeling	+0.10	-0.10
$t\bar{t}+light$ modeling	+0.06	-0.06
Luminosity	+0.03	-0.03
Light lepton (e, μ) id., isolation, trigger	+0.03	-0.03
Jet-vertex association, pileup modeling	+0.01	-0.01
Total systematic uncertainty	+0.64	-0.61
$t\bar{t}+ \geq 1b$ normalization	+0.12	-0.14
$t\bar{t}+ \geq 1c$ normalization	+0.03	-0.01
Statistical uncertainty	+0.21	-0.21
Total uncertainty	+0.68	-0.65

Table 7.14: Breakdown of the impact of uncertainties on signal strength. The background model statistics refers to the statistical uncertainties from limited number of simulated events and from estimated number of data events in the data-driven estimation of the non-prompt and fake lepton backgrounds component in the single lepton channel. The normalization of the $t\bar{t}+ \geq 1b$ and $t\bar{t}+ \geq 1c$ factors are not included in the total statistical component which is different from what it is reported in the text.

The Asimov fit is constructed in such a way that the nuisance parameters corresponding to the systematic uncertainties are all centered at zero and the normalization factor is centered at 1. Figure 7.44 shows the twenty most important systematic uncertainties, ranked by their impact on the signal strength error.

The first four leading nuisance parameters are all related to the $t\bar{t}+ \geq 1b$ background model and are constrained to at least 0.5σ . The leading nuisance parameter arises from comparing two different MC generators, where a different approximation for the multi parton final state predictions are varied simultaneously with the NLO generator, the number of partons in the matrix element, the parton shower, hadronization, and underlying event. This uncertainty has a post-fit contribution of $^{+0.45}_{-0.43}$ to the error on μ , and is constrained to 0.47σ .

The nuisance parameter related to the $t\bar{t}H$ signal model is ranked fifth in Table 7.44. The ($t\bar{t}H$: PS& hadronization) systematic accounts for the differences between PYTHIA 8 and HERWIG++ showering. However, its impact on the on signal sensitivity is still sub-dominant compared to the ones arising from the $t\bar{t}+ \geq 1b$ background model.

The majority of the parameters related to the detector performance are not constrained with respect to their prior uncertainties with few exceptions. The systematic uncertainty related to the first eigenvector in the light-jet efficiency uncertainty decomposition, referred to as *light*-tag Eigenvar 0, is constrained to 0.54σ , as shown in Figure 7.44. A high impact on the sensitivity from this nuisance parameter is not expected, since the definition of the most sensitive signal region requires at least four b -tagged jets using the tightest b -tagging operating point. This operating point has a very high rejection of *light*-jets, as mentioned in Section 7.5.3. However, the *light*-tag Eigenvar 0 nuisance parameter is found to be correlated to the nuisance parameters associated with $t\bar{t}$ +jets background model that have the largest impact on the sensitivity of the $t\bar{t}H(H \rightarrow b\bar{b})$ search.

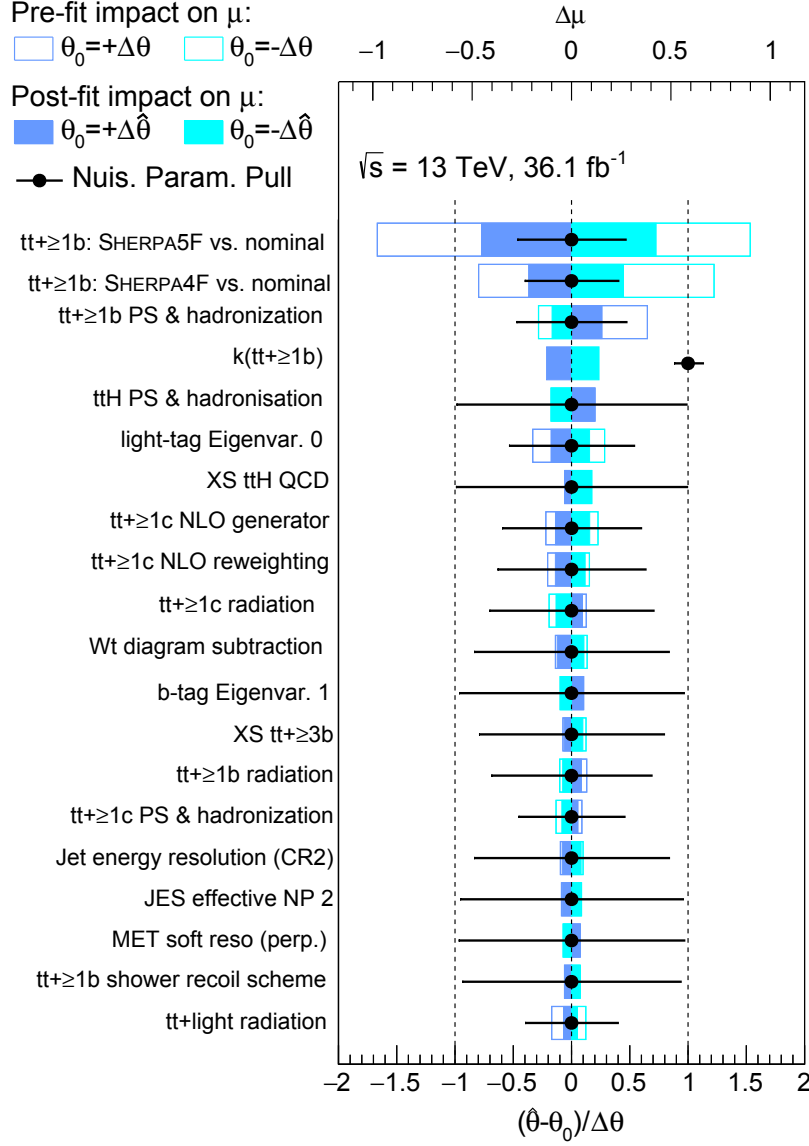


Figure 7.44: Ranking of the nuisance parameters included in the Asimov fit in the single lepton channel according to their impact on the measured signal strength μ . Only the top 20 parameters are shown. Nuisance parameters corresponding to MC statistical uncertainties are not considered here. The empty blue rectangles correspond to the pre-fit impact on μ and the filled blue ones to the post-fit impact on μ , both referring to the upper scale. The impact of each nuisance parameter, $\Delta\mu$, is computed by comparing the nominal best-fit μ with the result of the fit when fixing the considered nuisance parameter to its best-fit value, $\hat{\theta}$, shifted by its pre-fit (post-fit) uncertainties $\pm\Delta\theta$ ($\pm\Delta\hat{\theta}$). The black points show the pulls of the nuisance parameters with respect to their nominal values, θ_0 . These pulls and their relative post-fit errors, $\Delta\hat{\theta}/\Delta\theta$, refer to the lower scale. The parameter $k(t\bar{t} + \geq 1b)$ refers to the floating normalization of the $t\bar{t} + \geq 1b$ background, which is centered at 1 in the Asimov fit.

Pseudo Data Fits to Alternative $t\bar{t}$ Model

The robustness of the fit with respect to the choice of a particular $t\bar{t}$ +jets background model is verified by fits to pseudo-data built from an alternative $t\bar{t}$ model. The pseudo-data in this fit is built from the nominal predictions of all processes but the $t\bar{t}$ +jets background for which the POWHEG+PYTHIA8 is replaced by POWHEG+PYTHIA6. The POWHEG+PYTHIA6 is generated according to the settings used in the publication in Run 1 [152]. This sample was found to have sufficient MC statistics in order to have a meaningful fit result without large statistical fluctuations. A well understood fit is expected to use the nuisance parameters associated with $t\bar{t}$ +jets background to correct for the difference between POWHEG+PYTHIA8 and POWHEG+PYTHIA6, while leaving the other nuisance parameters and the signal strength untouched. In particular, the free-floating $t\bar{t}+ \geq 1b$ and $t\bar{t}+ \geq 1c$ normalization factors are expected to compensate for the difference between both setups.

A summary of the normalization factors ($k(t\bar{t}+ \geq 1b)$ and $k(t\bar{t}+ \geq 1c)$) and the measured signal strength is illustrated in Table 7.15 for the single-lepton channel. The normalization factors are found to be compatible with the ratio of the fractions in the POWHEG+PYTHIA 8 sample to POWHEG+PYTHIA 6 sample, 1.03 for $t\bar{t}+ \geq 1b$ and 0.87 for $t\bar{t}+ \geq 1c$. A difference of about 20% is found in the case of the normalization factor related to $t\bar{t}+ \geq 1c$. The effect from the MC statistics has been quantified on the POWHEG+PYTHIA 6 sample, from running toy experiments where the pseudo-data is allowed to change within the corresponding MC statistics. The uncertainty on the signal strength due to the MC statistics is about 22% in the single-lepton channel.

The best-fit value for the set of nuisance parameters in the fit to alternative pseudo-data in the single lepton channel is illustrated in Figure 7.45, where the shifts from the initial values of 0 and the errors are reported in units of the pre-fit uncertainty of the given parameter. The main pulls in the theoretical systematic uncertainties, illustrated in Figure 7.45 (a), are related to the $t\bar{t}$ modeling. In particular, significant pulls on the $t\bar{t}$ + light radiation and $t\bar{t}+ \geq 1c$ parton shower, are observed. This could be caused by the difference in the jet multiplicity in the POWHEG+PYTHIA 8 and POWHEG+PYTHIA 6 samples due to a different settings of the h_{hdamp} parameter. Also few instrumental systematics, such as b -tagging, JES, and *light*-tag Eigenvar 0, are slightly pulled to

Alternative $t\bar{t}$ pseudo-data fit	
Signal strength	
$\mu_{t\bar{t}H}$	$0.90^{+0.61}_{-0.58}$
Normalization factors	
$k(t\bar{t}+ \geq 1b)$	$0.98^{+0.13}_{-0.12}$
$k(t\bar{t}+ \geq 1c)$	$0.70^{+0.32}_{-0.28}$

Table 7.15: The obtained signal strength and the normalization factors, from the fit obtained from alternative $t\bar{t}$ pseudo-data fit built using POWHEG+PYTHIA6 in the single-lepton channel.

correct for the NP associated with the $t\bar{t}$ +jets background. The fit uses some NPs that have shape or normalization freedom to either converge or to adjust other mismodeling, which is most probable related to the $t\bar{t}+ \geq 1b$ model. The overall described fits enhances the confidence in the robustness of the signal extraction against the choice of the $t\bar{t}$ +jets model.

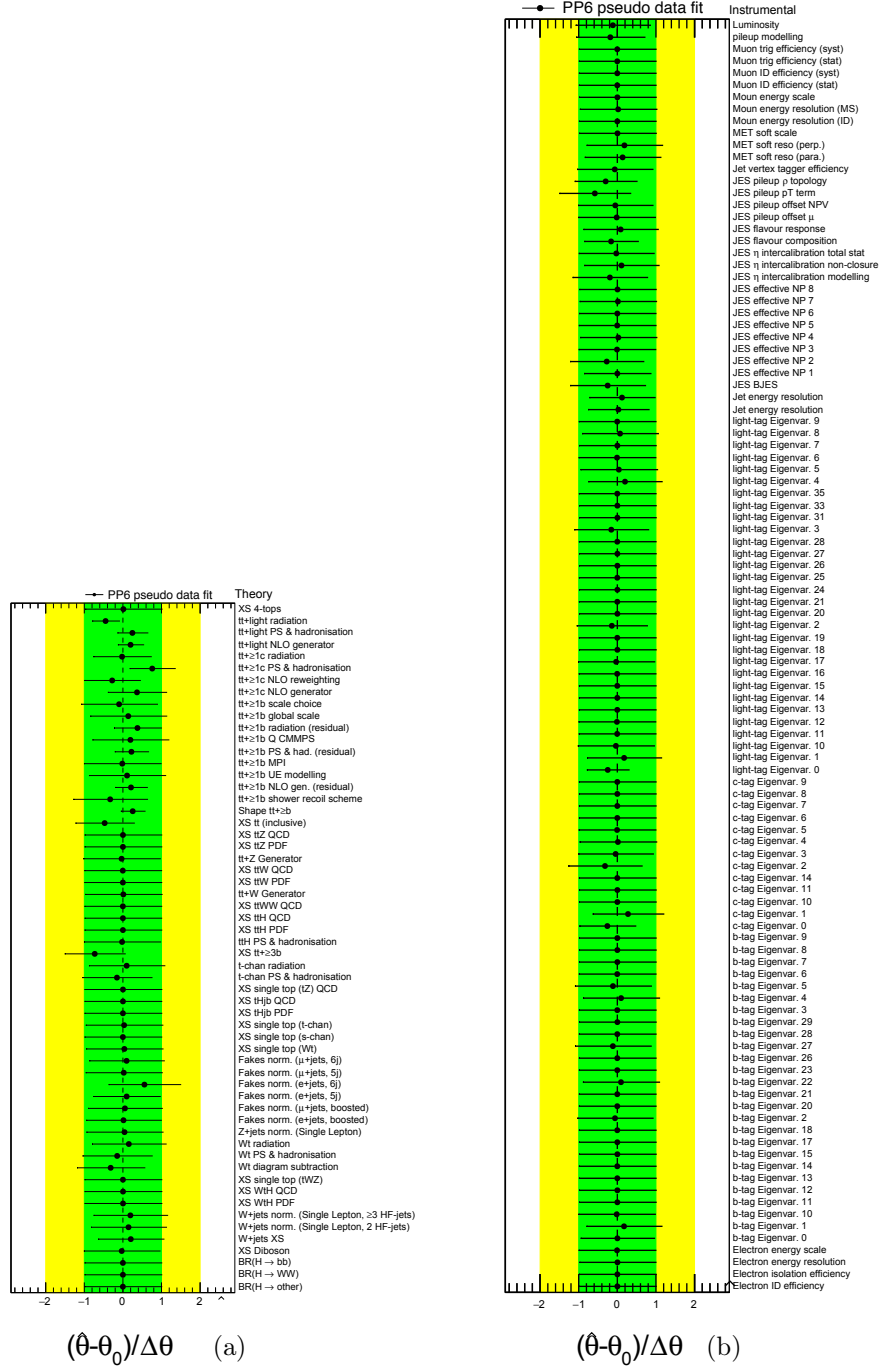


Figure 7.45: Post-fit (a) theoretical systematic uncertainties and (b) instrumental systematic uncertainties obtained from an Asimov fit to alternative pseudo-data built using POWHEG+PYTHIA 6, in the single lepton channel. The green (yellow) area represent the $\pm 1(2)\sigma$ band on the pre-fit systematic uncertainty. The position of the black points and the size of their horizontal bars represent the pulls and constraints in units of standard deviation, respectively.

7.10.4 Fit to Data and Results

After many checks, the fit was performed with the full measured data and the obtained results are described in the following. The best value of the signal strength for a Higgs boson mass of $m_H = 125 \text{ GeV}$ is:

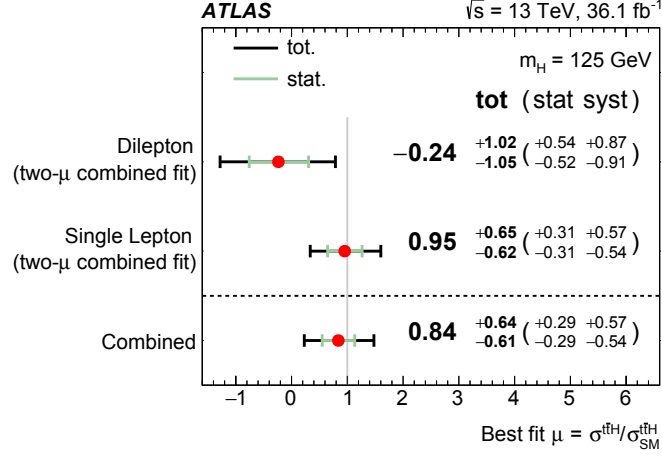
$$\mu = 0.84 \pm 0.29 \text{ (stat)}_{-0.54}^{+0.57} \text{ (syst)} = 0.84_{-0.61}^{+0.64}, \quad (7.13)$$

obtained from the combined fit in all signal and background regions in the two top decay channels. The statistical and systematic uncertainties are obtained as explained in the Asimov fit, described in Section 7.10.3.

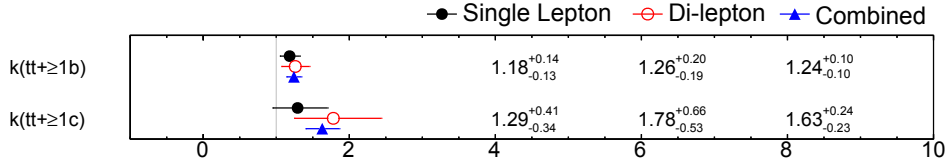
In order to consider the background constraint from the single and dilepton channels, an alternative combined fit was obtained. In this fit, an independent signal strengths were allowed for the single and dilepton channels, referred to as "two- μ " fit to data. The two- μ fit is a combined fit where nuisance parameters and k-factors are correlated between the two channels but considering the signal strengths in the single and dilepton channels as two separate normalization factors. The signal strength from this combined two- μ fit is given for each channel in Figure 7.46 (a). In the dilepton channel $\mu = -0.24_{-1.05}^{+1.02}$, and in the single-lepton channel $\mu = 0.95_{-0.62}^{+0.65}$. Note that a negative signal strength obtained in the two- μ combined fit in the dilepton channel means that the MC overestimates the background in the regions where the signal is expected. This can be caused by fluctuations of the MC samples. The measured signal strength from both fits are compatible within the uncertainties.

The normalization factors from the combined fit are found to be $1.24_{-0.10}^{+0.10}$ for $t\bar{t} + \geq 1b$, and $1.63_{-0.23}^{+0.24}$ for $t\bar{t} + \geq 1c$. The fit does very strongly constrain the free-floating background normalization factors as shown in Figure 7.46 (b). This is an improvement over the 8 TeV fiducial cross-section measurement for $t\bar{t}$ with one or two additional b -jets [151], where the error on the cross-section is about 30%. The obtained normalization factor for $t\bar{t} + \geq 1b$ has been checked with the on going $t\bar{t} + b\bar{b}$ measurement at 13 TeV and they are compatible within the assigned uncertainties.

Figure 7.47 shows the event yields observed in data compared to the prediction in each control and signal region, before the fit to data (pre-fit) and after the fit to data (post-fit), performed in all the analysis regions in the single-lepton and dilepton channels.



(a)



(b)

Figure 7.46: Summary of (a) the signal strength measurements, and (b) the normalization factors, in the individual channels and the combination. The numbers are obtained from a simultaneous fit in the two channels, but the measurements in the two channels separately are obtained keeping the signal strengths uncorrelated, while the nuisance parameters are kept correlated across channels

The normalization factors ($t\bar{t} + \geq 1b$, and $t\bar{t} + \geq 1c$) are set to 1 in the pre-fit plots which corresponds to considering the prediction from POWHEG+PYTHIA 8 for the fraction of each of these components with respect to the total $t\bar{t}$ prediction. In all analysis regions, the data agrees with the corrected prediction.

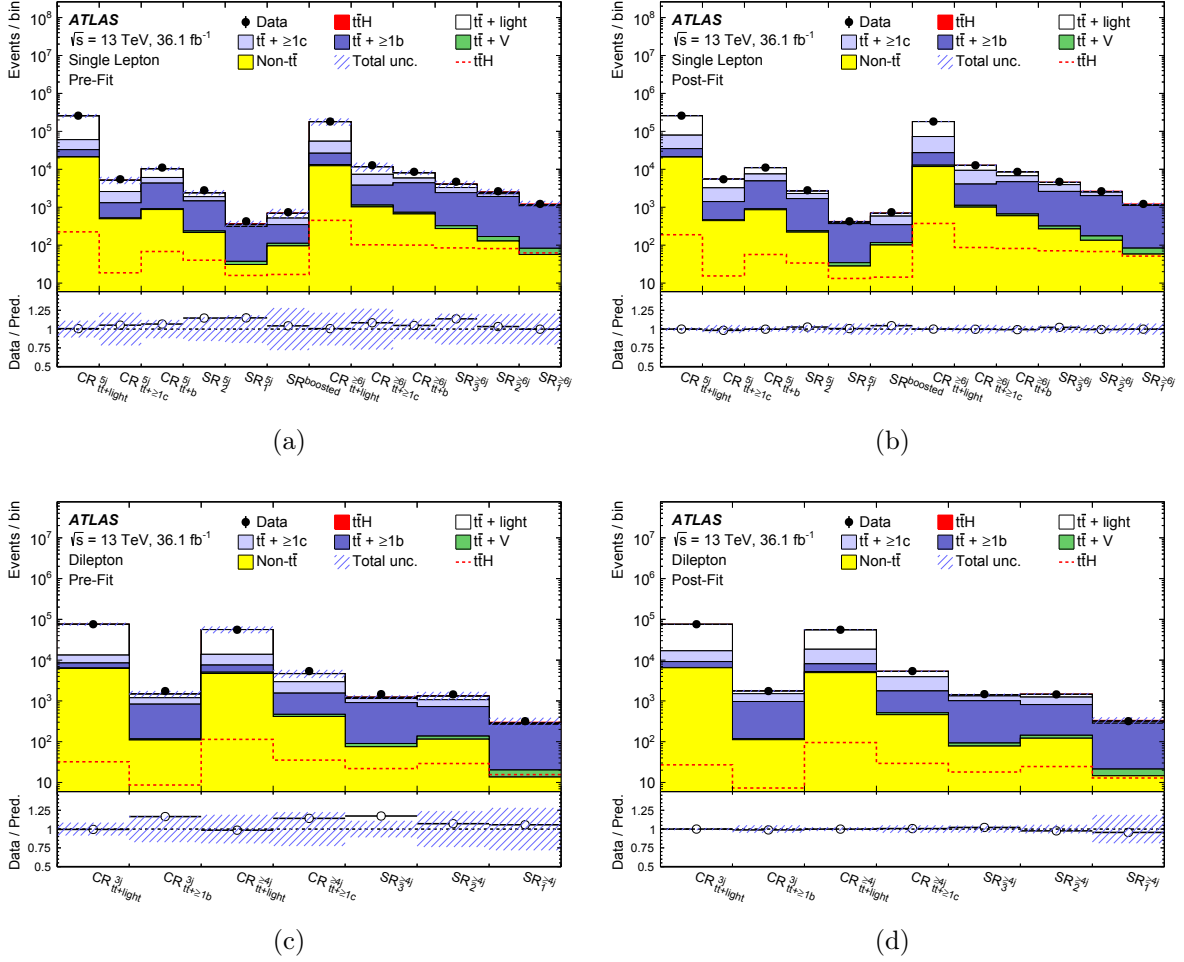


Figure 7.47: Comparison of predicted and observed event yields in all 19 analysis regions, (a-b) in the single lepton channel and (c-d) in the dilepton channel. (a) and (c) are pre-fit plots and (b) and (d) are the distributions after the combined dilepton and single lepton fit to data. The signal contribution is shown both as a filled red area stacked on top of the backgrounds and as a separate dashed red line. The hashed band represent the sum of the statistical and systematic uncertainties. Uncertainties on the $t\bar{t} + \geq 1b$ and $t\bar{t} + \geq 1c$ background normalizations are not included as those are free floating parameters of the fit. The pre-fit and post-fit yields for the single-lepton and the dilepton channels are summarized in Appendix A.4.

Figure 7.48 shows comparisons of the observed data and the prediction for the $H_{\text{T}}^{\text{had}}$ distribution in the $t\bar{t} + \geq 1c$ enriched control regions before and after applying the corrections from the fit (pre-fit and post-fit). The fit corrects the normalization mismodeling in these regions. The uncertainty is also reduced due to the constraints on the nuisance parameters.

Similarly, Figure 7.49 and 7.50 show comparisons for the observed data to the prediction of the classification BDT output in the five- and six-jet signal regions. The shape of the classification BDT output is modeled within the assigned uncertainties and the fit mainly corrects for the MC deficit normalization in several of these regions. Some fluctuations due to low statistics in some of the bins in the mentioned figures show a 1 or even 2σ deviations of data from MC predictions. However, the overall post-fit agreement is good. There is no clear trend of a shape mismodeling or normalization offset in the post-fit distributions, and the simultaneous fit of all bins is able to capture the mismodeling arising from $t\bar{t} + \text{jets}$ model.

The distributions in the dilepton channel can be found in Appendix A.5. Only variables with good agreement between data and MC simulation were considered as input variables to the classification BDT, which already leads to a good pre-fit agreement between the prediction and data in the signal regions.

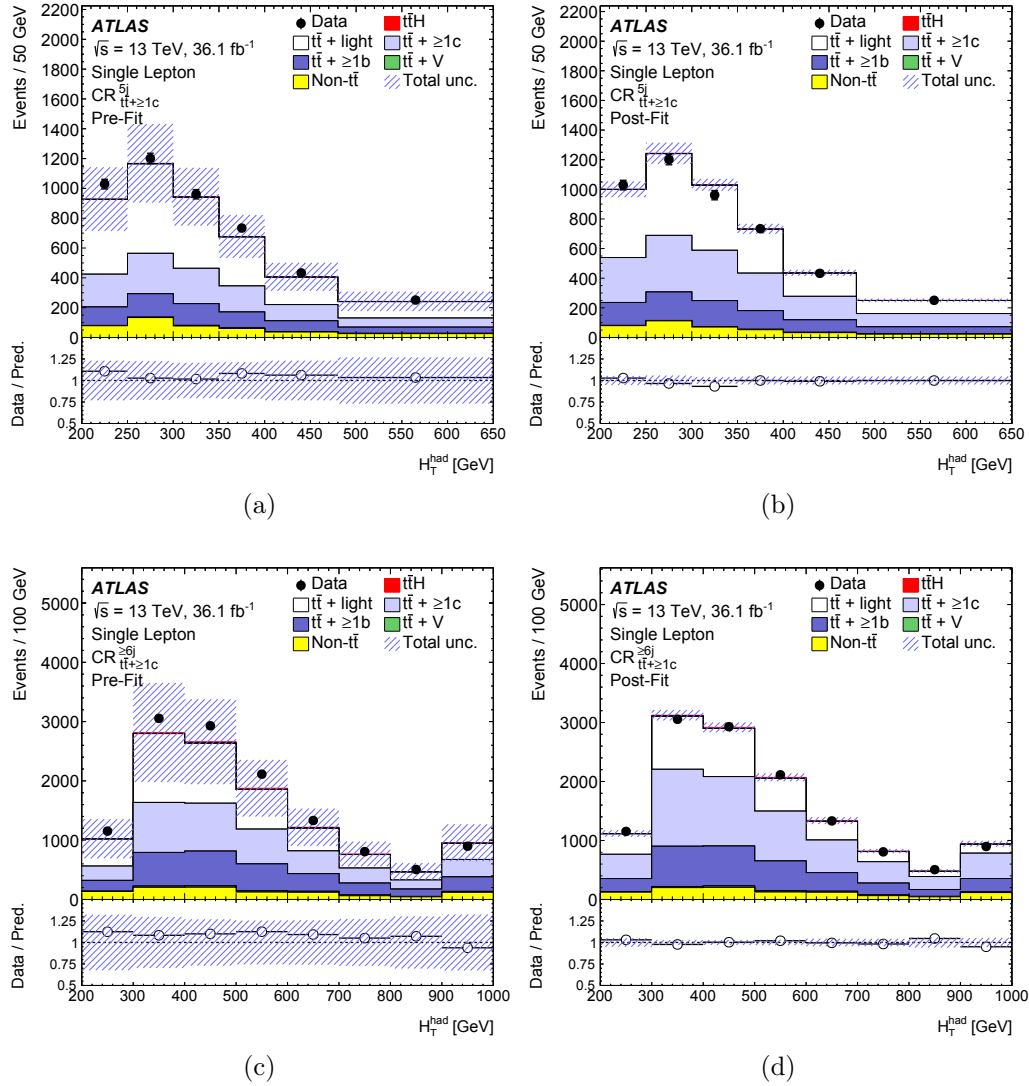


Figure 7.48: Comparison between data and prediction for the H_T^{had} distributions in the five- and six-jet $tt + \geq 1c$ enriched control regions. (a) and (c) are plots used as input to the fit, (b) and (d) are the distributions after the combined dilepton and single-lepton fit to data. The signal contribution is shown both as a filled red area stacked on top of the backgrounds and as a separate dashed red line. The hashed band represent the sum of the statistical and systematic uncertainties. Uncertainties on the $tt + \geq 1b$ and $tt + \geq 1c$ background normalizations are not included as those are free floating parameters of the fit.

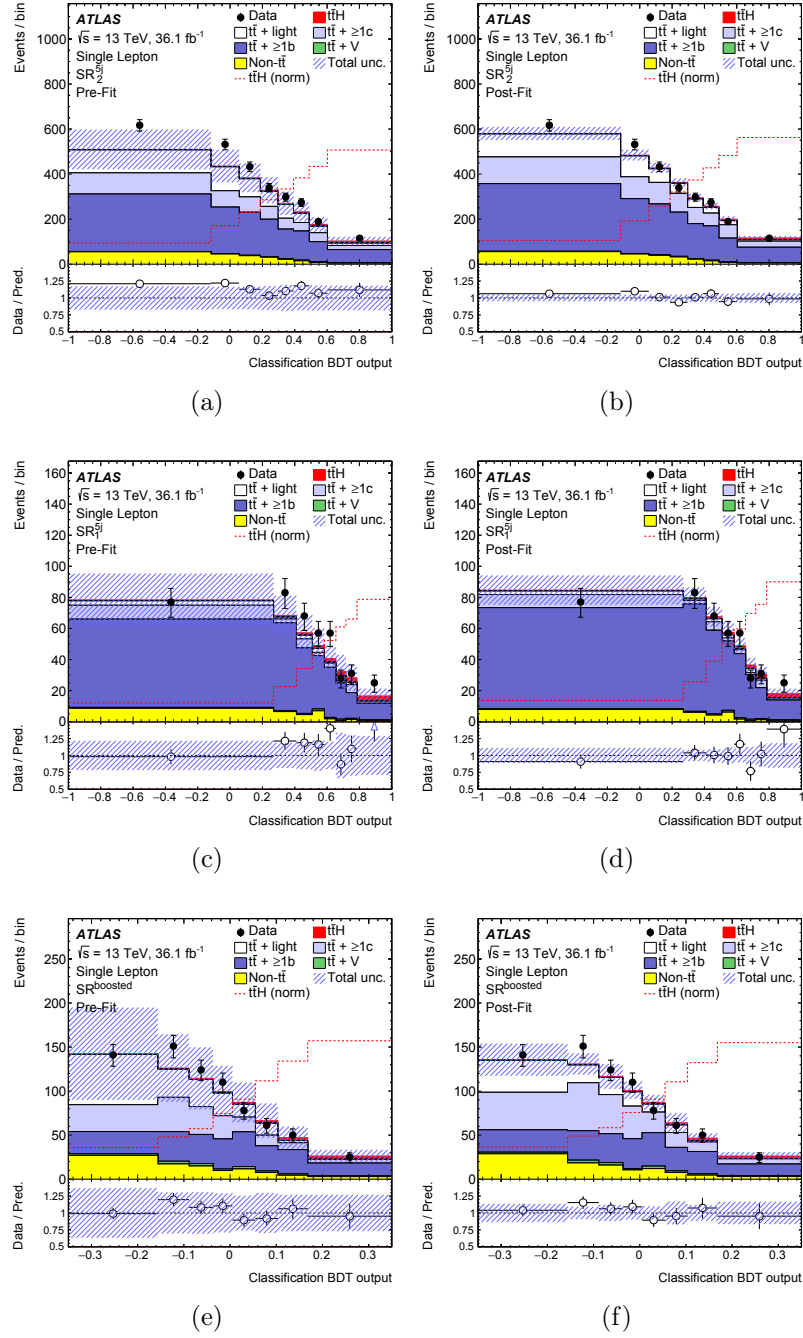


Figure 7.49: Comparison between data and prediction for the BDT output distributions in the five-jet and boosted signal regions. (a), (c), and (e) are plots used as input to the fit and (b), (d), and (f) are the distributions after the combined dilepton and single-lepton fit to data. The signal contribution is shown both as a filled red area stacked on top of the backgrounds and as a separate dashed red line. The hashed band represent the sum of the statistical and systematic uncertainties. Uncertainties on the $t\bar{t} + \geq 1b$ and $t\bar{t} + \geq 1c$ background normalizations are not included as those are free floating parameters of the fit.

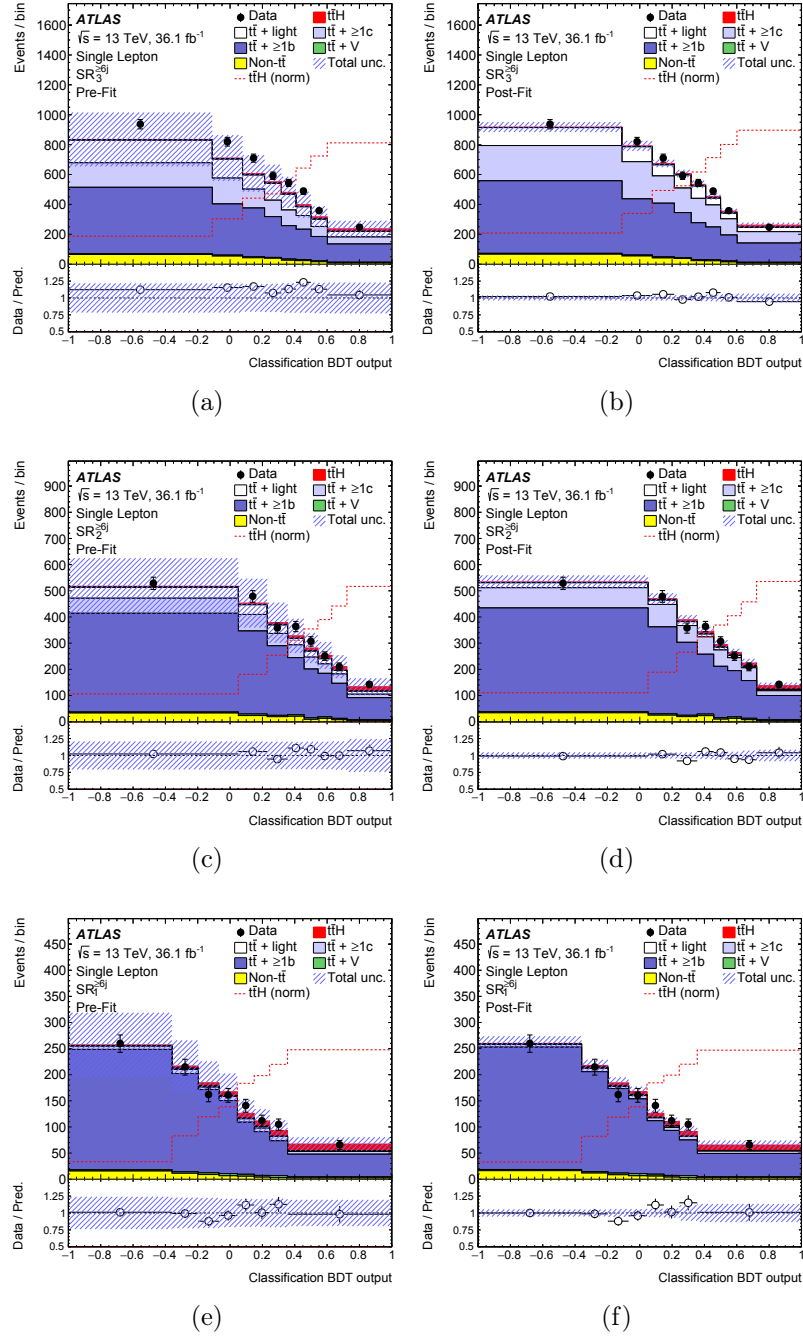


Figure 7.50: Comparison between data and prediction for the classification BDT output distributions in the six-jet signal regions. (a), (c), and (e) are plots used as input to the fit and (b), (d), and (f) are the distributions after the combined dilepton and single-lepton fit to data. The signal contribution is shown both as a filled red area stacked on top of the backgrounds and as a separate dashed red line. The hashed band represent the sum of the statistical and systematic uncertainties. Uncertainties on the $t\bar{t} + \geq 1b$ and $t\bar{t} + \geq 1c$ background normalizations are not included as those are free floating parameters of the fit.

The contributions from the different sources of uncertainty in the combined fit to μ are reported in Table 7.16. The total statistical uncertainty is evaluated, as described in the Asimov fit in Section 7.10.3. However, the contribution from the free-floating normalization factors is then included in the quoted total statistical uncertainty rather than in the systematic uncertainty component. The "intrinsic statistical uncertainty" refers to the statistical uncertainty evaluated after fixing the free-floating normalization factors. The total uncertainty is different from the sum in quadrature of the different components due to correlations between nuisance parameters built by the fit.

Uncertainty source	$\Delta\mu$	
$t\bar{t} + \geq 1b$ modeling	+0.46	-0.46
Background model statistics	+0.29	-0.31
b -tagging efficiency and mis-tag rates	+0.16	-0.16
Jet energy scale and resolution	+0.14	-0.14
$t\bar{t}H$ modeling	+0.22	-0.05
$t\bar{t} + \geq 1c$ modeling	+0.09	-0.11
JVT, pileup modeling	+0.03	-0.05
Other background modeling	+0.08	-0.08
$t\bar{t}$ +light modeling	+0.06	-0.03
Luminosity	+0.03	-0.02
Light lepton (e, μ) id., isolation, trigger	+0.03	-0.04
Total systematic uncertainty	+0.57	-0.54
$t\bar{t} + \geq 1b$ normalization	+0.09	-0.10
$t\bar{t} + \geq 1c$ normalization	+0.02	-0.03
Intrinsic statistical uncertainty	+0.21	-0.20
Total statistical uncertainty	+0.29	-0.29
Total uncertainty	+0.64	-0.61

Table 7.16: Breakdown of the impact of uncertainties on signal strength. The background model statistics refers to the statistical uncertainties from limited number of simulated events and from estimated number of data events in the data-driven estimation of fake lepton backgrounds component in the single lepton channel.

Figure 7.51 shows the twenty most important systematic uncertainties, ranked by their impact on the signal strength error. For each of these nuisance parameters, the best-fit value and the post-fit uncertainty are shown. Constraints on the data fit are very similar to the ones seen in the Asimov fit, as explained in Section 7.10.3.

Some nuisance parameters in Figure 7.51 are pulled in the fit from their nominal values. To understand the origin of these pulls, the corresponding nuisance parameters are

decorrelated across analysis regions and samples and the fit is repeated. These pulls are found to mainly correct the predictions of the $t\bar{t}$ background composition to the observed data in various regions. Similar pulls are also observed when a background-only fit is performed after removing the bins with the most significant signal contributions. The variations induced on the signal strength by the mentioned pulls are quantified by fixing the corresponding nuisance parameters to their pre-fit values, repeating the fit, and comparing the obtained μ value with the nominal one. These variations were found to be smaller than the uncertainty on the signal strength.

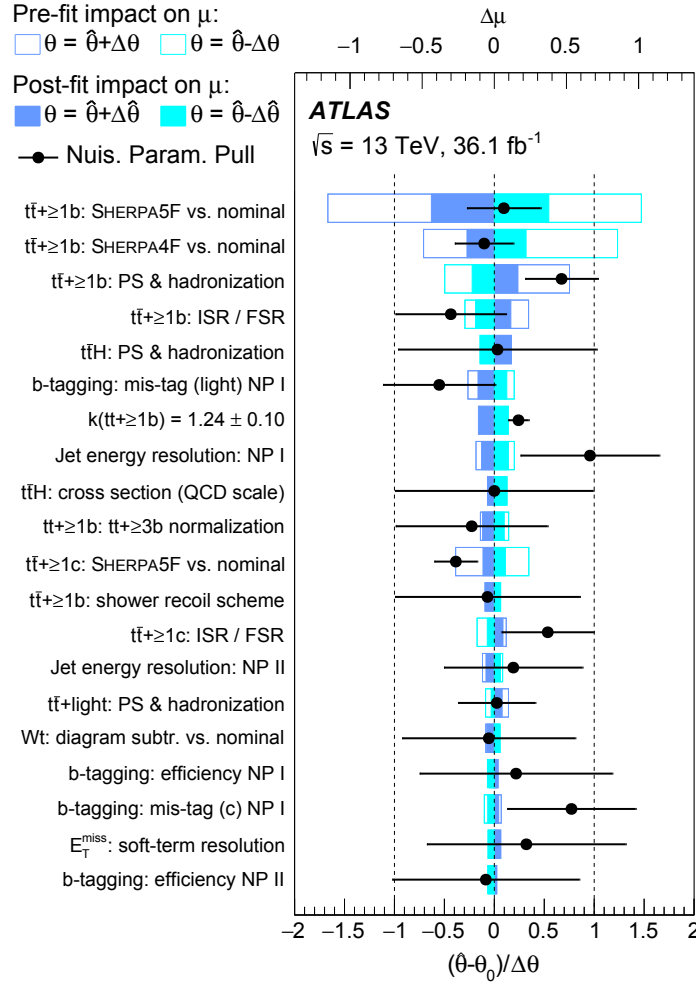


Figure 7.51: Ranking of the nuisance parameters included in the fit to data according to their impact on the measured signal strength μ . Only the top 20 parameters are shown. Nuisance parameters corresponding to MC statistical uncertainties are not considered here. The empty blue rectangles correspond to the pre-fit impact on μ and the filled blue ones to the post-fit impact on μ , both referring to the upper scale. The impact of each nuisance parameter, $\Delta\mu$, is computed by comparing the nominal best-fit μ with the result of the fit when fixing the considered nuisance parameter to its best-fit value, $\hat{\theta}$, shifted by its pre-fit (post-fit) uncertainties $\pm\Delta\theta$ ($\pm\Delta\hat{\theta}$). The black points show the pulls of the nuisance parameters with respect to their nominal values, θ_0 . These pulls and their relative post-fit errors, $\Delta\hat{\theta}/\Delta\theta$, refer to the lower scale. The parameter $k(t\bar{t} + \geq 1b)$ refers to the floating normalization of the $t\bar{t} + \geq 1b$ background, for which the pre-fit impact on μ is not defined, and for which both θ_0 and $\Delta\theta$ are set to 1. For experimental uncertainties which are broken down into several independent sources, the corresponding nuisance parameter (NP) index is reported.

Nuisance parameters are included in the maximum likelihood fit as uncorrelated parameters. However, the fit creates correlations between complementary nuisance parameters. Figure 7.52 shows the linear correlation coefficients of a selection of the systematic uncertainties reported in the ranking plot in Figure 7.51, obtained by the full fit to data. The most important correlations are the ones between the nuisance parameters and the signal strength ($\mu_{t\bar{t}H}$) which affect the sensitivity of the analysis. The most noticeable is the 66% anti-correlation between the " $t\bar{t}+ \geq 1b$: SHERPA5F vs. nominal" uncertainty and the uncertainty on the signal strength. This is due to the fact the most sensitive regions in the analysis are dominated by the background coming from $t\bar{t}+ \geq 1b$, and this correlation is therefore related to the difficulty to separate the signal from the background in these regions. This strong correlation together with the difference in MC modeling of the $t\bar{t}+ \geq 1b$ background has the dominant effect on the signal strength $\mu_{t\bar{t}H}$.

μ_{th}	100.0	-14.8	0.8	-66.0	-32.8	26.4	19.0	-13.0	-15.2	-11.4	-14.9	-8.0	7.6	-9.8	4.9	-8.3	5.1	6.9	3.0
k(tt+≥1b)	-14.8	100.0	-38.8	24.8	-22.0	37.2	7.5	3.1	5.7	-17.0	8.8	-9.3	15.6	9.8	27.8	7.0	6.4	38.9	9.8
k(tt+≥1c)	0.8	-38.8	100.0	-2.8	-6.7	-19.2	2.5	33.8	4.0	16.8	-17.5	16.7	-13.1	-22.7	18.9	-1.3	3.9	-10.6	9.9
tt+≥1b: SHERPA5F vs. nominal	-66.0	24.8	-2.8	100.0	16.4	-2.9	4.1	8.4	18.1	13.6	12.5	11.0	-17.6	-9.0	-5.6	5.4	-1.0	-10.7	-3.7
tt+≥1b: SHERPA4F vs. nominal	-32.8	-22.0	-6.7	16.4	100.0	-43.2	-12.6	-4.0	8.5	-4.9	2.4	-3.2	-2.3	-0.3	-8.8	4.2	1.6	-5.7	-0.7
tt+≥1b: PS & hadronisation	26.4	37.2	-19.2	-2.9	-43.2	100.0	29.9	-15.2	-16.3	24.2	-16.3	5.1	16.9	-9.3	-4.9	-7.2	-0.4	14.2	6.8
tt+≥1b: ISR / FSR	19.0	7.5	2.5	4.1	-12.6	29.9	100.0	4.7	-19.6	-6.8	-13.1	8.6	-15.7	-16.8	6.4	2.1	-5.6	-24.6	-7.0
b-tagging: mis-tag (light), NP 0	-13.0	3.1	33.8	8.4	-4.0	-15.2	4.7	100.0	-5.6	3.1	23.4	4.3	0.3	-14.5	4.2	1.6	-2.4	-9.8	-5.5
Jet energy resolution: NP 1	-15.2	5.7	4.0	18.1	8.5	-16.3	-19.6	-5.6	100.0	-4.5	15.3	0.6	-16.4	7.3	0.1	-8.8	-1.0	8.7	-3.3
tt+≥1b: tt+≥3b normalisation	-11.4	-17.0	16.8	13.6	-4.9	24.2	-6.8	3.1	-4.5	100.0	-5.4	-25.8	1.5	-7.0	-3.6	-4.6	4.7	-2.7	8.2
tt+≥1c: SHERPA5F vs. nominal	-14.9	8.8	-17.5	12.5	2.4	-16.3	-13.1	23.4	15.3	-5.4	100.0	3.1	-25.2	19.7	13.9	-4.2	-1.5	-3.8	-6.5
tt+≥1b: shower recoil scheme	-8.0	-9.3	16.7	11.0	-3.2	5.1	8.6	4.3	0.6	-25.8	3.1	100.0	-8.0	1.7	-4.6	-1.4	4.9	-12.7	8.8
tt+≥1c: ISR / FSR	7.6	15.6	-13.1	-17.6	-2.3	16.9	-15.7	0.3	-16.4	1.5	-25.2	-8.0	100.0	-11.3	6.9	0.5	-1.3	25.4	1.5
Jet energy resolution: NP 0	-9.8	9.8	-22.7	-9.0	-0.3	-9.3	-16.8	-14.5	7.3	-7.0	19.7	1.7	-11.3	100.0	24.8	-8.4	-3.1	-10.8	-9.3
tt+light: PS & hadronisation	4.9	27.8	18.9	-5.6	-8.8	-4.9	6.4	4.2	0.1	-3.6	13.9	-4.6	6.9	24.8	100.0	-4.6	6.6	7.5	5.3
Wt: diagram subtr. vs. nominal	-8.3	7.0	-1.3	5.4	4.2	-7.2	2.1	1.6	-8.8	-4.6	-4.2	-1.4	0.5	-8.4	-4.6	100.0	-0.1	8.0	-2.1
b-tagging: efficiency, NP 1	5.1	6.4	3.9	-1.0	1.6	-0.4	-5.6	-2.4	-1.0	4.7	-1.5	4.9	-1.3	-3.1	6.6	-0.1	100.0	2.4	-7.5
b-tagging: mis-tag (c), NP 0	6.9	38.9	-10.6	-10.7	-5.7	14.2	-24.6	-9.8	8.7	-2.7	-3.8	-12.7	25.4	-10.8	7.5	8.0	2.4	100.0	5.4
b-tagging: efficiency, NP 0	3.0	9.8	9.9	-3.7	-0.7	6.8	-7.0	-5.5	-3.3	8.2	-6.5	8.8	1.5	-9.3	5.3	-2.1	-7.5	5.4	100.0

Figure 7.52: Linear correlation coefficients for the signal strength and the nuisance parameters obtained from the combined fit to data. Only a selection of the systematic uncertainties reported in the ranking plot in Figure 7.51, are shown. The upper and lower triangles are symmetric by construction.

7.10.5 Setting Limits

If the measured signal strength shows no significant excess with respect to the background-only hypothesis, an upper limit can be set on the production cross section for the $t\bar{t}H$ process by performing hypothesis tests based on a frequentist approach.

The test statistic q_μ for this hypothesis test is defined as the profile likelihood ratio [202],

$$q_\mu(x) = -2\ln\left(\frac{\mathcal{L}(x|\mu, \hat{\theta}_\mu)}{\mathcal{L}(x|\hat{\mu}, \hat{\theta})}\right), \quad (7.14)$$

where \mathcal{L} is the likelihood function of the profile likelihood, and x stands for the data or pseudo data. The $\hat{\mu}$ and $\hat{\theta}$ are the parameters that maximise the likelihood (with the constraint $0 \leq \hat{\mu} \leq \mu$), and $\hat{\theta}_\mu$ are the values of the nuisance parameters that maximise the likelihood function for a given value of μ .

The test statistic, defined in Equation 7.14, is used to evaluate the validity of the background-only hypothesis with $\mu = 0$, and to make statistical inferences about μ , such as upper limits using CL_s method [202–204] as implemented in the RooFIT package [199,200]. The ingredients of the CL_s method are shown in Figure 7.53 which represent a hypothetical example of the distributions of the test statistic for the two hypothesis; the background only hypotheses $f(q_\mu|b)$ in the right distribution and the signal plus background hypothesis $f(q_\mu|s+b)$ in the left distribution. The compatibility among the observed data (q_{obs}) and a given hypothesis is measured by a p-value:

$$p_{s+b} = P(q \leq q_{obs}|s+b) = \int_{q_{obs}}^{\infty} f(q_\mu|s+b) dq_\mu. \quad (7.15)$$

$$1 - p_b = P(q \leq q_{obs}|b) = \int_{q_{obs}}^{\infty} f(q_\mu|b) dq_\mu. \quad (7.16)$$

Using the above two variables, the CL_s variables is defined as:

$$CL_s(\mu) = \frac{p_{s+b}}{1 - p_b}. \quad (7.17)$$

The 95% Confidence Level upper limit on μ , referred to as $\mu^{95\%CL}$, is the value of μ for which $CL_s = 0.05$. Therefore, a value of μ above $\mu^{95\%CL}$ is excluded at 95% confidence level.

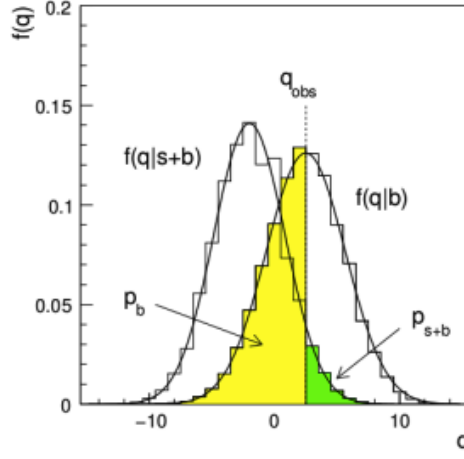


Figure 7.53: Example of the distribution of the test statistics for background-only and signal+background hypothesis.

No significant excess is observed in data compared to the background-only hypothesis. The uncertainty bands ($\pm 1\sigma$, and $\pm 2\sigma$) on the median for the background-only hypothesis are evaluated from the crossing of the cumulative probability distribution with the corresponding quantiles without recalculation of the probability distribution function. Figure 7.54 shows the observed and expected 95% confidence level upper limits on the signal strength. The combined fit finds a 1.4σ excess of $t\bar{t}H(H \rightarrow b\bar{b})$ over the background only hypothesis. A signal strength higher than 2.0 is excluded at the 95% confidence level, compared to an expected exclusion limit of 1.2 in the absence of signal.

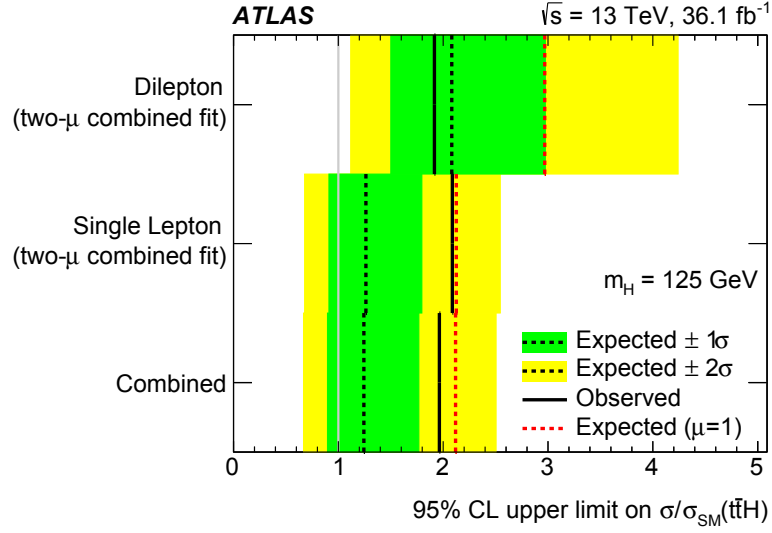


Figure 7.54: Summary of the 95% confidence level upper limits on the $\sigma(t\bar{t}H)$ relative to the SM prediction in the individual channels and for the combination. The observed limits are shown, together with the expected limits both in the background-only hypothesis (dotted black lines) and the SM hypothesis (dotted red lines). In the case of the expected limits in the background-only hypothesis, one- and two- standard-deviation uncertainty bands are also shown. The numbers of the expected and observed upper limits are summarized in Appendix A.6.

Chapter 8

CONCLUSION AND OUTLOOK

The discovery of the Higgs boson was an important triumph for the LHC. It opened a new sector of measurements dedicated to understanding the properties of this newly discovered particle. So far, all measurements are compatible with the hypothesis of a Standard Model Higgs boson. However, several properties are still missing, such as the Higgs boson coupling to quarks.

The coupling of the Higgs boson to the top-quark is expected to be the largest in the Standard Model, since the top-quark is by far the heaviest known particle. The production of a Standard Model Higgs boson in association with a pair of top-quarks, $t\bar{t}H$, is a channel that allows a direct measurement of this coupling at the LHC.

This thesis presented a search for the production of $t\bar{t}H$, in the $H \rightarrow b\bar{b}$ channel, using a dataset corresponding to an integrated luminosity of 36.1 fb^{-1} of proton-proton collisions at $\sqrt{s} = 13 \text{ TeV}$ recorded by the ATLAS experiment at the LHC. Events with one or two charged leptons from the decay of the top-quark pair are considered. Events are split into non-overlapping regions based on the number of jets and the number of b -tagged jets in order to provide regions enhanced in the signal and the dominant background components. Multivariate techniques based on Boosted Decision Trees are used to discriminate between the $t\bar{t}H$ signal and the dominant background originating from $t\bar{t}$ +jets. The modeling of additional b -tagged jets is crucial for this search, and methods were developed to constrain this large background with the latest theoretical predictions. Misidentified fake and non-prompt lepton backgrounds, are estimated using data-driven techniques based on the Matrix Method. All analysis regions are combined into a profile likelihood fit to test for the presence of signal, which simultaneously determines the event yields for the signal while constraining the overall background model within the assigned systematic uncertainties.

A combined signal strength of $0.84^{+0.64}_{-0.61}$ is observed, corresponding to a 1.4σ excess of $t\bar{t}H(H \rightarrow b\bar{b})$ over the background hypothesis in data. This result excluded $t\bar{t}H(H \rightarrow b\bar{b})$ cross-sections two times larger than the SM prediction at a 95% confidence level. While this has still large uncertainties, there is a 60% improvement in sensitivity with respect to the analysis performed with 20.3 fb^{-1} of data at $\sqrt{s} = 8 \text{ TeV}$.

In addition to the results presented in this thesis, the ATLAS collaboration performed searches for the $t\bar{t}H$ production in various decay modes of the Higgs boson. The additional searches are categorized in two channels: $H \rightarrow \gamma\gamma$, and $H \rightarrow \text{multilepton}$. The later includes the Higgs boson decays to WW^* , $\tau\tau$, and ZZ^* . The combination of the results presented in this thesis with the other $t\bar{t}H$ searches from the ATLAS experiment has an observed significance of 4.2σ , compared to an expectation of 3.8σ [184]. This provides the first experimental evidence for the $t\bar{t}H$ production mode.

Future measurements using more data will benefit greatly from reducing the dominant systematic uncertainties on the reconstructed signal, primarily from the components associated with the MC model for the $t\bar{t} + \geq 1b$ background. Therefore, the most important aspect for any increase in the sensitivity is to design the analysis to be less dependent on the MC used for the estimation of the dominant background and to improve the modeling of the $t\bar{t} + \geq 1b$ background. Measurements of the $t\bar{t} + b\bar{b}$ process at 13 TeV will be an important input for improving the MC generators. Moreover, efforts to merge the NLO $t\bar{t} + b\bar{b}$ calculation with the inclusive $t\bar{t}$ production is essential to profit from the most precise calculation while keeping a large phase space to constrain and control the background predictions. The second largest impact on the sensitivity of the signal strength is the statistical uncertainty on the MC predictions. Efforts to increase the amount of generated Monte Carlo events in the phase space relevant to $t\bar{t}H$ are crucial for future measurements.

The estimate of the fake and non-prompt leptons using the Matrix Method, works well in the control regions, but improvements will be needed in the future to avoid the large statistical fluctuations in the signal regions, where fakes and non-prompt leptons are highly suppressed due to the high b -jet multiplicities. For example, this could be achieved by triggering on events with looser identification and isolation requirements with high jet multiplicities.

The LHC and the ATLAS experiment will continue to collect data at 13 and 14 TeV in the coming years. A 5σ observation of the $t\bar{t}H$ production is expected with the full 100 fb^{-1} of Run 2 data, expected by the end of 2018. This will provide a first direct measurement of the top Higgs Yukawa coupling. However, a complete understanding the top Higgs Yukawa coupling requires a determination of the sign of this coupling. This can be measured through the production of the Higgs boson in association with a single

top-quark (tH and tWH), which has an even smaller cross section than the $t\bar{t}H$ process, requiring more data for a 5σ observation. The era of Higgs physics is only beginning, and the top Higgs Yukawa couplings along with other properties will be a major focus of collider physics for the coming years.

Appendix A

ADDITIONAL MATERIAL

A.1 The Boosted Category

In the boosted category, jets reconstructed with the default algorithm (small-R jets) are used as inputs for further re-clustering [205] through an anti- k_t algorithm with a radius parameter of $R = 1.0$ resulting in a collection of the so called *large-R* jets. An event in the boosted $t\bar{t}H$ category is categorized as such if it contains:

- One large-R jet with $p_T > 200$ GeV and two b -tagged small-R jets, which represents the boosted Higgs boson candidate. The b -tagging requirement corresponds to the 85% working point.
- One large-R jet with $p_T > 250$ GeV and exactly one b -tagged small-R jet that represent the boosted top-quark candidate.
- An additional small-R jet which is b -tagged at 85% and it is not one of the sub-jets of either the Higgs boson or the hadronic top-quark candidate that represent the b -jet from the leptonic top-quark.

A cartoon illustrating the object selection of the boosted region is shown in Figure A.1.

A.2 Region Definition in the dilepton channel

Three signal regions are defined in the dilepton channel where the highest $t\bar{t}H$ signal purity, referred to as $SR_1^{\geq 4j}$, is defined by requiring at least four jets among which three are b -tagged using the very-tight operating point at 60% and the other is b -tagged using the tight operating point at 70%. The other two signal regions are defined with looser b -tagging requirements, and referred to as $SR_2^{\geq 4j}$, and $SR_3^{\geq 3j}$. The remaining dilepton events that have at least four jets are classified into two control regions, one of which is enriched in $t\bar{t} + \text{light}$, $CR_{t\bar{t} + \text{light}}^{\geq 4j}$, and the second one in $t\bar{t} + \geq 1c$, $CR_{t\bar{t} + \geq 1c}^{\geq 4j}$. Moreover, events with three jets are split into two control regions enriched in $t\bar{t} + \text{light}$ and $t\bar{t} + \geq 1b$, $CR_{t\bar{t} + \text{light}}^{3j}$ and $CR_{t\bar{t} + \geq 1b}^{3j}$, respectively. Figure A.2 details the definition of the three-jet

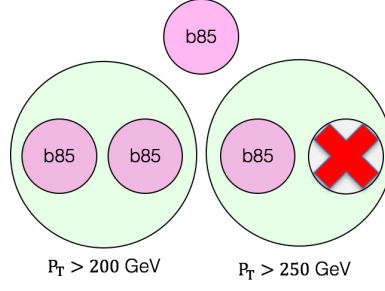


Figure A.1: Cartoon illustrating the boosted region object selection. The Higgs boson candidate is indicated by the left green circle representing the large-R jet with $p_T > 200 \text{ GeV}$ and two b -tagged small-R jets (purple circles). The hadronic top-quark candidate is indicated by the right green circle that has one large-R jet with $p_T > 250 \text{ GeV}$ and exactly one b -tagged at 85% small-R jet and one light jet labelled by the red "X". The additional small-R jet which is b -tagged at 85% and it is not one of the sub-jets of either the Higgs boson or the hadronic top-quark candidate reclustered jets is indicated by the small purple circle on the top.

and four-jet signal and control regions for the dilepton channel depending on the b -tagging requirements.

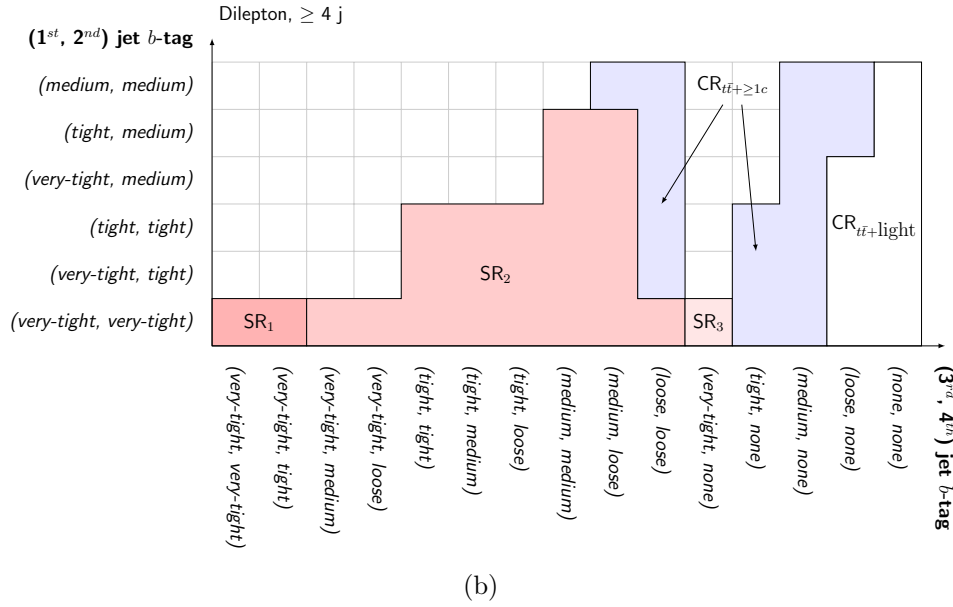
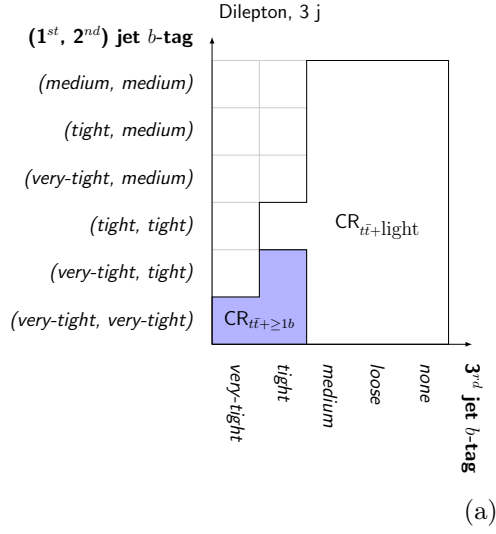


Figure A.2: Definition of the (a) three-jet and (b) four-jet signal and control regions in the dilepton channel, in terms of b -tagging requirements. The vertical axis shows the requirements on the first two jets, while the horizontal axis on the third jet and/ or fourth jets. The jets are ordered such that the ones passing tighter b -tagging requirements are considered first, which means the empty squares are not physical.

A.3 $t\bar{t}$ +HF modeling

Several normalized distributions of various variables in the $t\bar{t} + b$ category are shown in Figures [A.3](#), and [A.4](#), and in the $t\bar{t} + \geq 3b$ category in Figures [A.5](#) and [A.6](#).

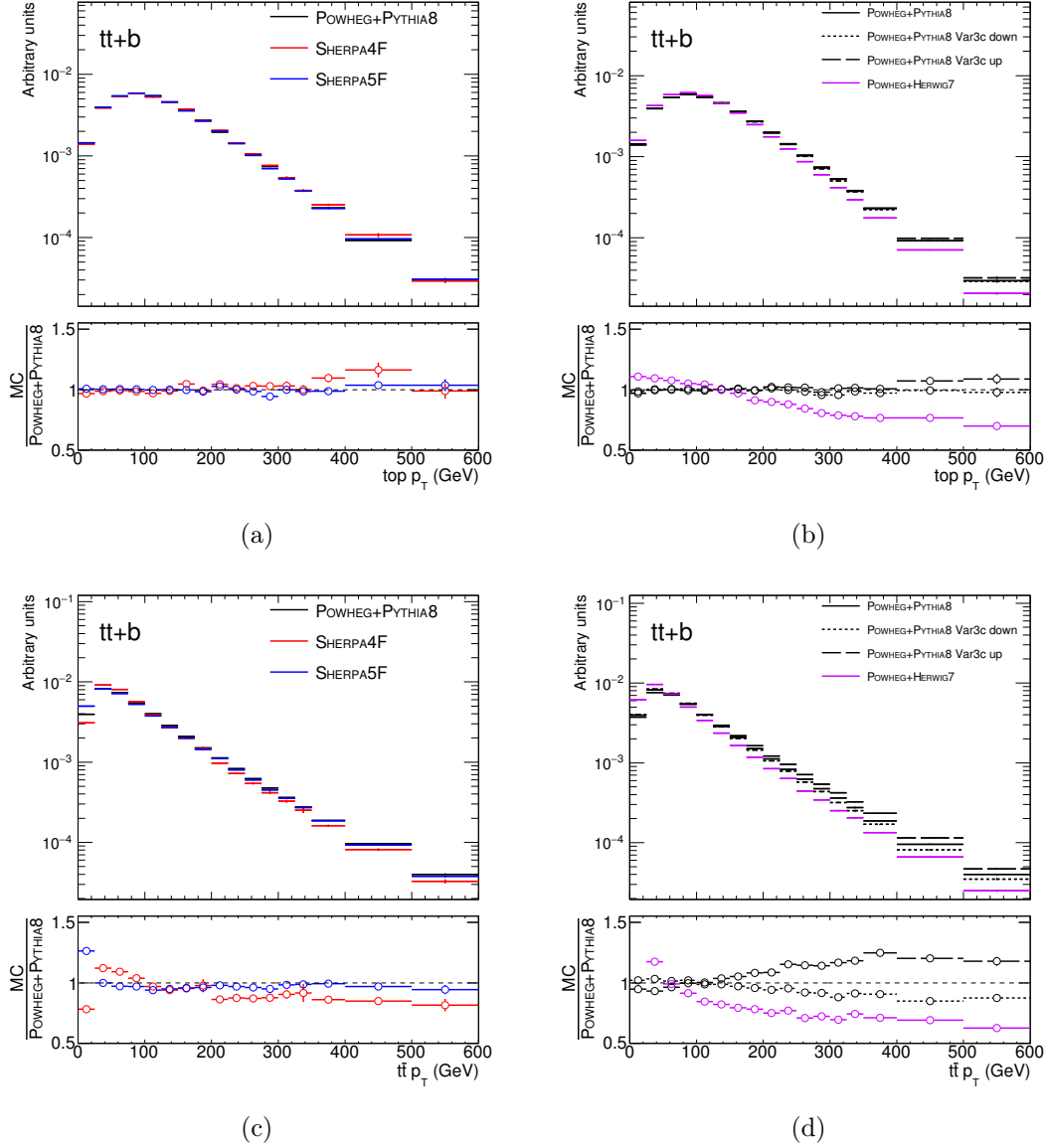


Figure A.3: Comparison of normalized kinematic variables in the $t\bar{t} + b$ category: (a) and (b) show top-quark transverse momentum (p_T^{top}), (c) and (d) show the transverse momentum of the $t\bar{t}$ system, ($p_T^{t\bar{t}}$). (a) and (c) show the differences among the 5F and 4F scheme by comparing POWHEG+PYTHIA 8 and SHERPA4F. (b) and (d) show the differences among the nominal POWHEG+PYTHIA 8 and the $t\bar{t}$ alternative samples: the impact of factorization and renormalization scale variations, and the radiation systematics for POWHEG-BOX+PYTHIA 8 sample (dashed black line), parton shower systematic using POWHEG+HERWIG 7 (purple line), and generator systematic using SHERPA5F (blue line).

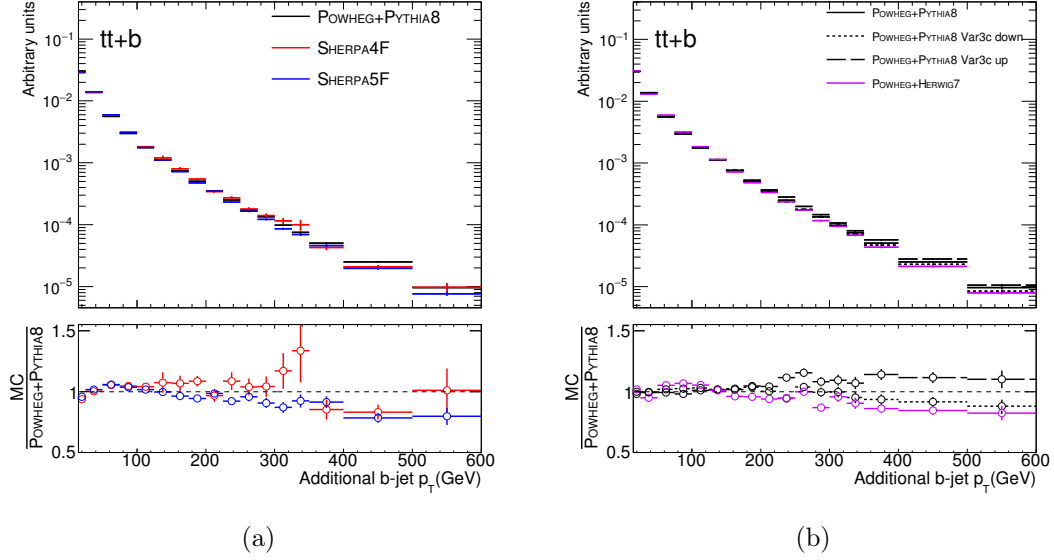


Figure A.4: Comparison of the normalized transverse momentum of the additional b -jet (p_T^b) that does not originate from the decay of the $t\bar{t}$ system, in the $t\bar{t} + b$ category. (a) shows the differences among the 5F and 4F scheme by comparing POWHEG+PYTHIA 8 and SHERPA4F. (b) shows the differences among the nominal POWHEG+PYTHIA 8 and the $t\bar{t}$ alternative samples: the impact of factorization and renormalization scale variations, and the radiation systematics for POWHEG-BOX+PYTHIA 8 sample (dashed black line), parton shower systematic using POWHEG+HERWIG 7 (purple line), and generator systematic using SHERPA5F (blue line).

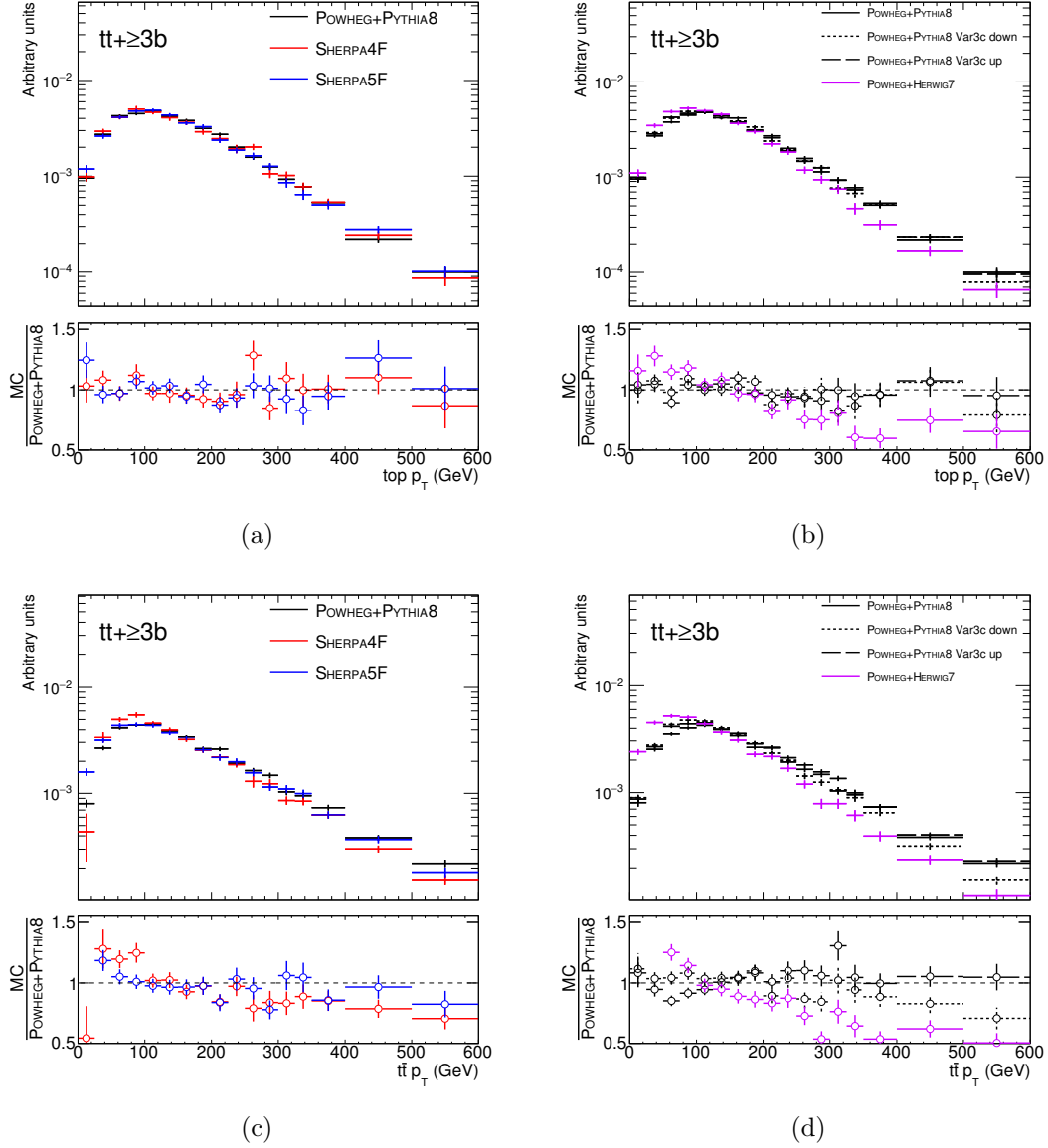


Figure A.5: Comparison of normalized kinematic variables in the $t\bar{t} + \geq 3b$ category: (a) and (b) show top-quark transverse momentum (p_T^{top}), (c) and (d) show the transverse momentum of the $t\bar{t}$ system, ($p_T^{t\bar{t}}$). (a) and (c) show the differences among the 5F and 4F scheme by comparing POWHEG+PYTHIA 8 and SHERPA4F. (b) and (d) show the differences among the nominal POWHEG+PYTHIA 8 and the $t\bar{t}$ alternative samples: the impact of factorization and renormalization scale variations, and the radiation systematics for POWHEG-BOX+PYTHIA 8 sample (dashed black line), parton shower systematic using POWHEG+HERWIG 7 (purple line), and generator systematic using SHERPA5F (blue line).

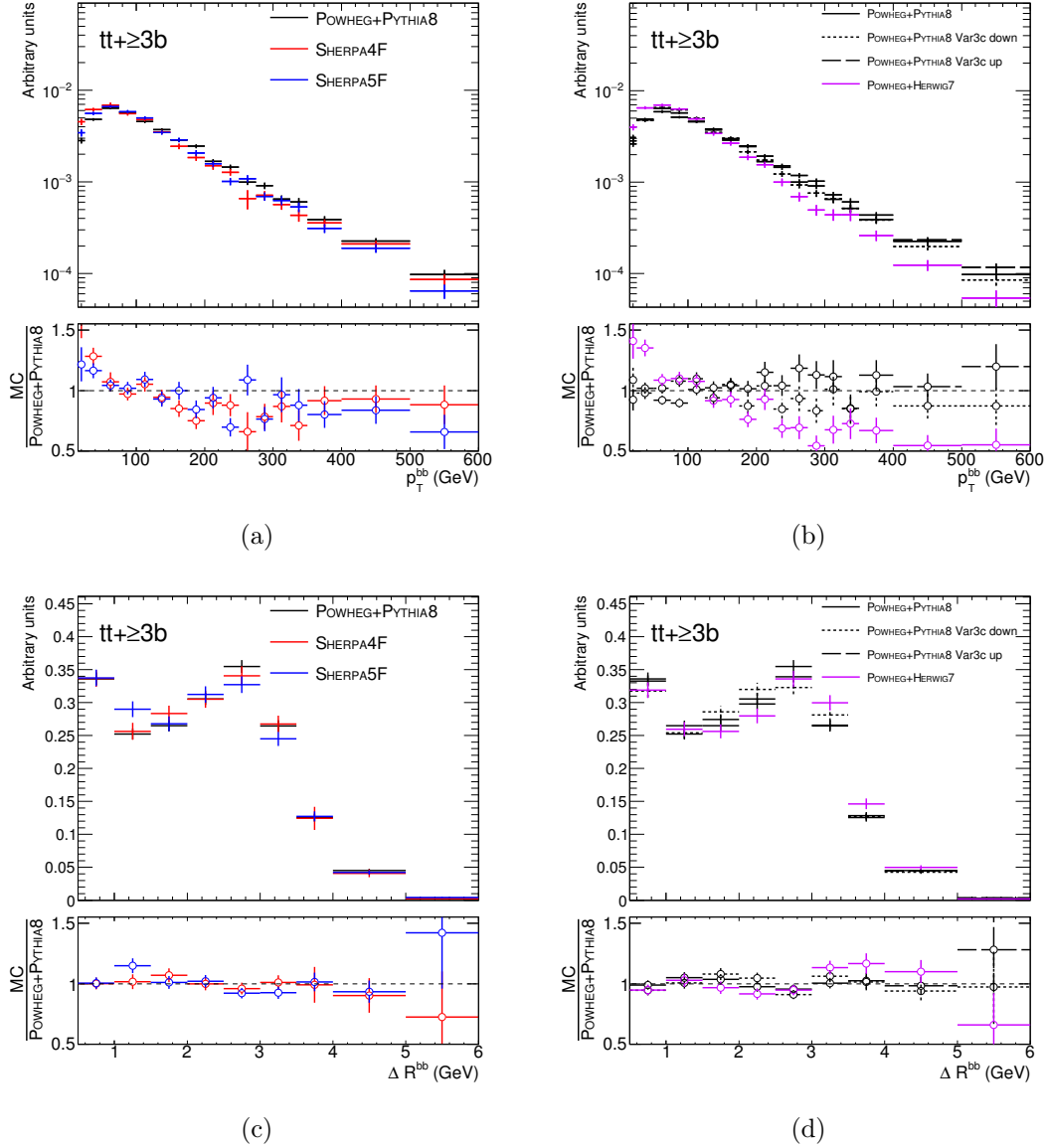


Figure A.6: Comparison of normalized kinematic variables in the $t\bar{t} + \geq 3b$ category: (a) and (b) show the transverse momentum of the two additional b -jets (p_T^{bb}) that do not originate from the decay of the $t\bar{t}$ system, (c) and (d) show the opening angle between the two additional (b -jets ΔR^{bb}). (a) and (c) show the differences among the 5F and 4F scheme by comparing POWHEG+PYTHIA 8 and SHERPA4F. (b) and (d) show the differences among the nominal POWHEG+PYTHIA 8 and the $t\bar{t}$ alternative samples: the impact of factorization and renormalization scale variations, and the radiation systematics for POWHEG-BOX+PYTHIA 8 sample (dashed black line), parton shower systematic using POWHEG+HERWIG 7 (purple line), and generator systematic using SHERPA5F (blue line).

A.4 Event Yields

The pre-fit and post-fit yields for the single-lepton channel are summarized in Table A.1, and A.2 and the ones for dilepton channel are in Table A.3. The uncertainties on the post-fit yields are computed with the following approximation $err = \sqrt{\sum_{sys}^i \sum_{sys}^j c_{ij} \times \Delta y_i \Delta y_j}$, where c_{ij} is the correlation coefficient and Δy_i is the effect of the yields due to the variation of the i^{th} nuisance parameters within its error. A noticeable reduction of systematics is observed in the post-fit yields compared to pre-fit ones.

(a)

Sample	$CR_{t\bar{t}+light}^{5j}$		$CR_{t\bar{t}+\geq 1c}^{5j}$		$CR_{t\bar{t}+b}^{5j}$	
	Pre-fit	Post-fit	Pre-fit	Post-fit	Pre-fit	Post-fit
$t\bar{t}H$	224 \pm 22	190 \pm 140	18.7 \pm 2.5	15 \pm 12	68.0 \pm 7.6	57 \pm 42
$t\bar{t} + light$	197 000 \pm 26 000	179 900 \pm 4900	2580 \pm 720	2300 \pm 210	4250 \pm 920	3560 \pm 240
$t\bar{t} + \geq 1c$	27 500 \pm 4300	44 100 \pm 5500	1280 \pm 500	1840 \pm 250	1770 \pm 270	2590 \pm 390
$t\bar{t} + \geq 1b$	11 300 \pm 1100	13 500 \pm 1300	790 \pm 130	944 \pm 94	3400 \pm 440	4030 \pm 320
$t\bar{t} + V$	589 \pm 55	584 \pm 54	23.2 \pm 4.1	21.3 \pm 2.9	48.1 \pm 5.9	46.6 \pm 5.4
Non- $t\bar{t}$	21 300 \pm 4100	20 900 \pm 3200	520 \pm 180	440 \pm 100	960 \pm 190	860 \pm 160
Total	258 000 \pm 29 000	259 320 \pm 910	5200 \pm 1100	5560 \pm 160	10 400 \pm 1300	11 140 \pm 290
Data	259320		5465		11095	

(b)

Sample	SR_2^{5j}		SR_1^{5j}		$SR^{boosted}$	
	Pre-fit	Post-fit	Pre-fit	Post-fit	Pre-fit	Post-fit
$t\bar{t}H$	40.1 \pm 5.1	34 \pm 25	15.9 \pm 2.1	13.3 \pm 9.8	16.9 \pm 1.9	14 \pm 10
$t\bar{t} + light$	500 \pm 210	393 \pm 67	15 \pm 33	12.5 \pm 9.3	180 \pm 120	112 \pm 32
$t\bar{t} + \geq 1c$	436 \pm 92	610 \pm 100	30 \pm 17	28 \pm 14	168 \pm 70	235 \pm 39
$t\bar{t} + \geq 1b$	1230 \pm 200	1450 \pm 110	273 \pm 53	335 \pm 25	236 \pm 89	229 \pm 33
$t\bar{t} + V$	19.9 \pm 2.9	19.7 \pm 2.4	6.4 \pm 1.3	6.4 \pm 1.2	16.1 \pm 2.9	16.6 \pm 2.4
Non- $t\bar{t}$	269 \pm 64	220 \pm 52	54 \pm 11	28.1 \pm 8.4	104 \pm 30	101 \pm 26
Total	2440 \pm 390	2724 \pm 70	371 \pm 68	423 \pm 23	710 \pm 200	708 \pm 40
Data	2798		426		740	

Table A.1: Comparison of predicted and observed event yields in each of the five-jet control and signal regions, and the boosted signal region, in the single-lepton channel.

(a)

Sample	$\text{CR}_{t\bar{t}+\text{light}}^{\geq 6j}$		$\text{CR}_{t\bar{t}+\geq 1c}^{\geq 6j}$		$\text{CR}_{t\bar{t}+b}^{\geq 6j}$	
	Pre-fit	Post-fit	Pre-fit	Post-fit	Pre-fit	Post-fit
$t\bar{t}H$	450 \pm 48	370 \pm 280	102 \pm 13	87 \pm 64	100 \pm 12	83 \pm 61
$t\bar{t} + \text{light}$	125 000 \pm 34 000	108 200 \pm 4300	4300 \pm 2000	3350 \pm 430	2220 \pm 520	1820 \pm 170
$t\bar{t} + \geq 1c$	28 400 \pm 7200	45 700 \pm 5100	3600 \pm 1300	5300 \pm 680	1460 \pm 330	2080 \pm 300
$t\bar{t} + \geq 1b$	13 100 \pm 1800	14 600 \pm 1400	2660 \pm 540	2950 \pm 280	3670 \pm 500	4080 \pm 320
$t\bar{t} + V$	1010 \pm 120	996 \pm 91	118 \pm 21	118 \pm 14	70.5 \pm 8.5	67.9 \pm 7.2
Non- $t\bar{t}$	12 600 \pm 3000	11 800 \pm 2000	1060 \pm 340	1000 \pm 210	710 \pm 160	600 \pm 110
Total	181 000 \pm 39 000	181 690 \pm 860	11 800 \pm 3200	12 810 \pm 260	8200 \pm 1100	8730 \pm 230
Data	181706		12778		8576	

(b)

Sample	$\text{SR}_3^{\geq 6j}$		$\text{SR}_2^{\geq 6j}$		$\text{SR}_1^{\geq 6j}$	
	Pre-fit	Post-fit	Pre-fit	Post-fit	Pre-fit	Post-fit
$t\bar{t}H$	85 \pm 10	71 \pm 52	81 \pm 10	68 \pm 50	62 \pm 11	51 \pm 38
$t\bar{t} + \text{light}$	750 \pm 370	586 \pm 98	210 \pm 210	96 \pm 33	14 \pm 10	12.1 \pm 5.8
$t\bar{t} + \geq 1c$	880 \pm 350	1330 \pm 190	350 \pm 100	473 \pm 99	53 \pm 33	44 \pm 20
$t\bar{t} + \geq 1b$	2100 \pm 420	2290 \pm 170	1750 \pm 370	1850 \pm 130	1010 \pm 240	1032 \pm 59
$t\bar{t} + V$	51.2 \pm 7.4	50.8 \pm 5.9	40.8 \pm 5.7	40.3 \pm 4.8	25.8 \pm 3.7	25.3 \pm 3.2
Non- $t\bar{t}$	303 \pm 82	267 \pm 63	155 \pm 52	134 \pm 46	75 \pm 20	58 \pm 17
Total	4140 \pm 850	4590 \pm 110	2550 \pm 510	2657 \pm 82	1220 \pm 250	1223 \pm 42
Data	4698		2641		1222	

Table A.2: Comparison of predicted and observed event yields in each of the six-jet control and signal regions, in the single-lepton channel.

(a)

Sample	$\text{CR}_{t\bar{t}+\text{light}}^{3j}$		$\text{CR}_{t\bar{t}+\geq 1b}^{3j}$		$\text{CR}_{t\bar{t}+\text{light}}^{\geq 4j}$		$\text{CR}_{t\bar{t}+\geq 1c}^{\geq 4j}$	
	Pre-fit	Post-fit	Pre-fit	Post-fit	Pre-fit	Post-fit	Pre-fit	Post-fit
$t\bar{t}H$	32.2 \pm 3.8	27 \pm 20	8.7 \pm 1.1	7.3 \pm 5.4	114 \pm 11	95 \pm 70	35.3 \pm 3.6	29 \pm 22
$t\bar{t} + \text{light}$	63 100 \pm 5500	59 100 \pm 1400	290 \pm 110	255 \pm 44	42 500 \pm 9700	37 100 \pm 1300	1730 \pm 730	1410 \pm 180
$t\bar{t} + \geq 1c$	4800 \pm 2100	7700 \pm 1100	360 \pm 160	536 \pm 89	6300 \pm 2800	10 300 \pm 1400	1410 \pm 590	2160 \pm 290
$t\bar{t} + \geq 1b$	2130 \pm 230	2620 \pm 240	710 \pm 140	848 \pm 75	2510 \pm 280	2850 \pm 290	1080 \pm 120	1240 \pm 110
$t\bar{t} + V$	113 \pm 31	112 \pm 29	7 \pm 27	7 \pm 30	350 \pm 180	330 \pm 170	52 \pm 41	50 \pm 39
Non- $t\bar{t}$	6300 \pm 1500	6500 \pm 1200	110 \pm 29	112 \pm 23	4700 \pm 1100	4930 \pm 910	420 \pm 120	460 \pm 100
Total	76 400 \pm 6500	76 010 \pm 390	1500 \pm 260	1765 \pm 60	56 000 \pm 11 000	55 650 \pm 420	4700 \pm 1100	5350 \pm 120
Data	76025		1744		55627		5389	

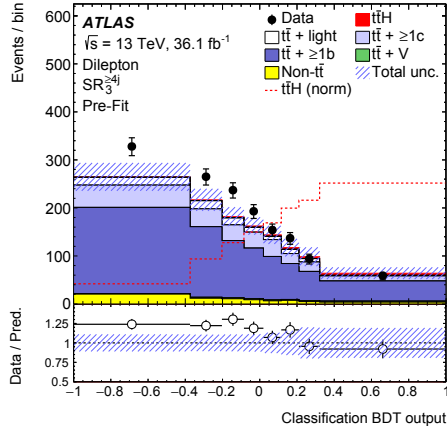
(b)

Sample	$\text{SR}_3^{\geq 4j}$		$\text{SR}_2^{\geq 4j}$		$\text{SR}_1^{\geq 4j}$	
	Pre-fit	Post-fit	Pre-fit	Post-fit	Pre-fit	Post-fit
$t\bar{t}H$	21.9 \pm 2.5	18 \pm 13	29.1 \pm 4.2	25 \pm 18	15.6 \pm 2.5	12.9 \pm 9.5
$t\bar{t} + \text{light}$	83 \pm 41	95 \pm 30	250 \pm 110	215 \pm 43	6.4 \pm 9.9	11.1 \pm 9.3
$t\bar{t} + \geq 1c$	235 \pm 61	313 \pm 53	340 \pm 210	427 \pm 89	12.6 \pm 9.4	25.8 \pm 7.8
$t\bar{t} + \geq 1b$	819 \pm 85	917 \pm 71	590 \pm 96	669 \pm 59	247 \pm 61	263 \pm 20
$t\bar{t} + V$	15 \pm 35	15 \pm 34	22 \pm 38	22 \pm 39	7 \pm 56	7 \pm 57
Non- $t\bar{t}$	75 \pm 17	78 \pm 16	115 \pm 36	121 \pm 29	13.6 \pm 3.8	14.6 \pm 3.8
Total	1250 \pm 140	1436 \pm 55	1350 \pm 320	1479 \pm 66	302 \pm 85	334 \pm 59
Data	1467		1444		319	

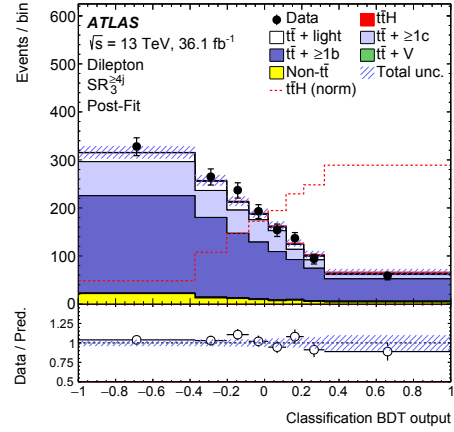
Table A.3: Comparison of predicted and observed event yields in the dilepton channel (a) control regions and (b) signal regions. Post-fit yields are after the combined fit in dilepton and single-lepton channels to data.

A.5 Additional plots

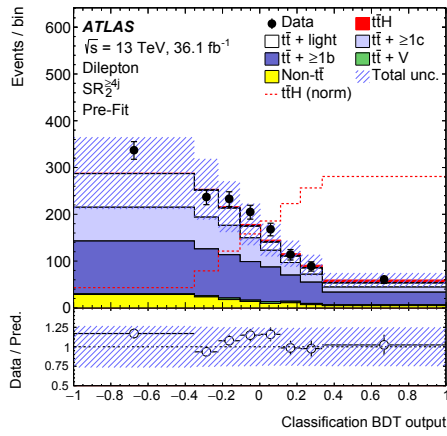
The distribution of the classification BDT output before and after the fit to data are shown in Figure [A.7](#) for the signal regions in the dilepton channel.



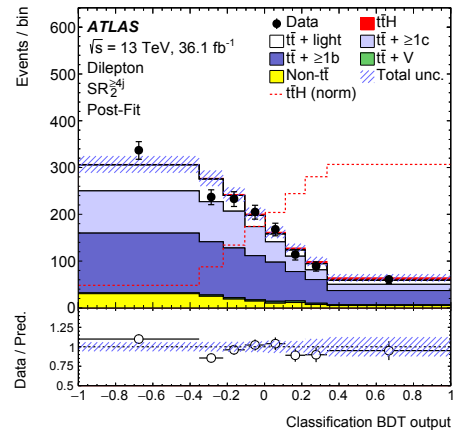
(a)



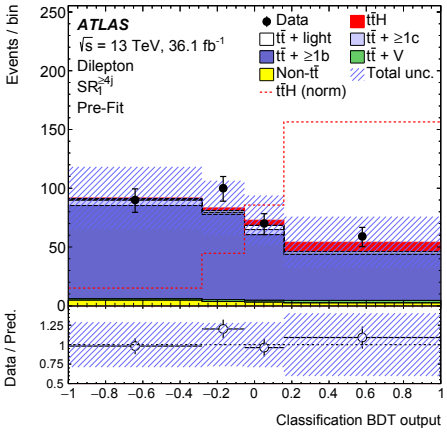
(b)



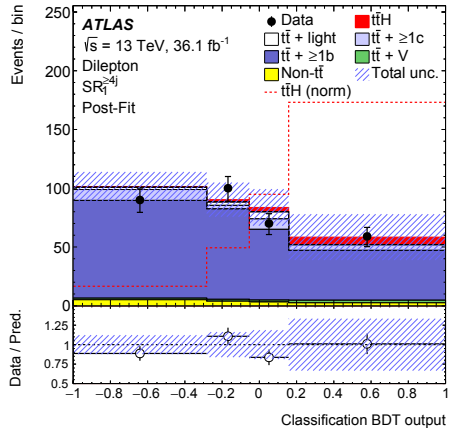
(c)



(d)



(e)



(f)

Figure A.7: Comparison between data and prediction for the classification BDT output distributions in the signal regions in the dilepton channel. (a), (c), and (e) are pre-fit plots and (b), (d), and (f) are the distributions after the combined dilepton and single-lepton fit to data. The pre-fit plots do not include an uncertainty on the $t\bar{t} + \geq 1b$ or $t\bar{t} + \geq 1c$ normalizations.

A.6 Setting limits

The observed and expected median upper limits on μ with a CL of 95% for the background-only hypothesis with $\mu = 0$ and the SM hypothesis $\mu = 1$ of a SM Higgs boson are summarized in Table A.4.

	Observed	Expected ($\mu = 0$)			Expected ($\mu = 1$)
		Median	+/-1 σ	+/-2 σ	
Dilepton	2.64	2.74	[1.98, 3.86]	[1.47, 5.43]	3.63
Single lepton	1.95	1.40	[1.01, 1.99]	[0.75, 2.82]	2.27
Dilepton (from two- μ fit)	1.84	2.47	[1.78, 3.48]	[1.32, 4.91]	3.39
Single Lepton (from two- μ fit)	2.09	1.26	[0.91, 1.80]	[0.68, 2.54]	2.13
Combined	1.96	1.24	[0.89, 1.77]	[0.67, 2.50]	2.12

Table A.4: Summary of the observed and expected upper limits on the $\sigma/\sigma_{\text{SM}}(t\bar{t}H)$ relative to the SM prediction of the 95% confidence level for the single-lepton, dilepton, and the combination obtained from the test statistic. The 69% and 95% confidence intervals around the expected limits under the background-only hypothesis are also provided, by the $\pm 1\sigma$ and $\pm 2\sigma$, respectively.

References

- [1] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, *Phys. Lett. B* **716** (2012) 1, [arXiv:1207.7214].
- [2] CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, *Phys. Lett. B* **716** (2012) 30, [arXiv:1207.7235].
- [3] S. L. Glashow, *Partial Symmetries of Weak Interactions*, *Nucl. Phys.* **22** (1961) 579–588.
- [4] S. Weinberg, *A Model of Leptons*, *Phys. Rev. Lett.* **19** (1967) 1264–1266.
- [5] G. 't Hooft and M. J. G. Veltman, *Regularization and Renormalization of Gauge Fields*, *Nucl. Phys.* **B44** (1972) 189–213.
- [6] F. Englert and R. Brout, *Broken Symmetry and the Mass of Gauge Vector Mesons*, *Phys. Rev. Lett.* **13** (1964) 321–323.
- [7] P. W. Higgs, *Broken symmetries, massless particles and gauge fields*, *Phys. Lett.* **12** (1964) 132–133.
- [8] P. W. Higgs, *Broken Symmetries and the Masses of Gauge Bosons*, *Phys. Rev. Lett.* **13** (1964) 508–509.
- [9] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble, *Global Conservation Laws and Massless Particles*, *Phys. Rev. Lett.* **13** (1964) 585–587.
- [10] P. W. Higgs, *Spontaneous Symmetry Breakdown without Massless Bosons*, *Phys. Rev.* **145** (1966) 1156–1163.
- [11] ATLAS Collaboration, *Measurements of the Higgs boson production and decay rates and coupling strengths using pp collision data at $\sqrt{s} = 7$ and 8 TeV in the ATLAS experiment*, *Eur. Phys. J. C* **76** (2016) 6, [arXiv:1507.0454].
- [12] CMS Collaboration, *Precise determination of the mass of the Higgs boson and tests of compatibility of its couplings with the standard model predictions using proton collisions at 7 and 8 TeV*, *Eur. Phys. J. C* **75** (2015) 212, [arXiv:1412.8662].
- [13] ATLAS and CMS Collaborations, *Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments*, *Phys. Rev. Lett.* **114** (2015) 191803, [arXiv:1503.0758].
- [14] ATLAS Collaboration, *Study of the spin and parity of the Higgs boson in diboson decays with the ATLAS detector*, *Eur. Phys. J. C* **75** (2015) 476, [arXiv:1506.0566].

- [15] CMS Collaboration, *Constraints on the spin-parity and anomalous HVV couplings of the Higgs boson in proton collisions at 7 and 8 TeV*, *Phys. Rev. D* **92** (2015) 012004, [[arXiv:1411.3441](#)].
- [16] C. Englert, A. Freitas, M. M. Mühlleitner, T. Plehn, M. Rauch, M. Spira, and K. Walz, *Precision Measurements of Higgs Couplings: Implications for New Physics Scales*, *J. Phys.* **G41** (2014) 113001, [[arXiv:1403.7191](#)].
- [17] ATLAS and CMS Collaborations, *Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV*, *JHEP* **08** (2016) 045, [[arXiv:1606.0226](#)].
- [18] LHC Higgs Cross Section Working Group Collaboration, *Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector*, [arXiv:1610.0792](#).
- [19] A. Salam, *Weak and Electromagnetic Interactions*, *Conf. Proc.* **C680519** (1968) 367–377.
- [20] Particle Data Group Collaboration, *Review of Particle Physics*, *Chin. Phys.* **C40** (2016), no. 10 100001.
- [21] E. Noether, *Invariant Variation Problems*, *Gott. Nachr.* **1918** (1918) 235–257, [[physics/0503066](#)]. [Transp. Theory Statist. Phys.1,186(1971)].
- [22] G. Rajasekaran, *Fermi and the Theory of Weak Interactions*, *Resonance J. Sci. Educ.* **19** (2014), no. 1 18–44, [[arXiv:1403.3309](#)].
- [23] J. Goldstone, A. Salam, and S. Weinberg, *Broken Symmetries*, *Phys. Rev.* **127** (1962) 965–970.
- [24] M. Kobayashi and T. Maskawa, *CP Violation in the Renormalizable Theory of Weak Interaction*, *Prog. Theor. Phys.* **49** (1973) 652–657.
- [25] ATLAS and CMS Collaborations, *Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV*, *JHEP* **08** (2016) 045, [[arXiv:1606.0226](#)].
- [26] ATLAS Collaboration, *Evidence for the Higgs-boson Yukawa coupling to tau leptons with the ATLAS detector*, *JHEP* **04** (2015) 117, [[arXiv:1501.0494](#)].
- [27] CMS Collaboration, *Evidence for the 125 GeV Higgs boson decaying to a pair of τ leptons*, *JHEP* **05** (2014) 104, [[arXiv:1401.5041](#)].
- [28] ATLAS Collaboration, *Evidence for the $H \rightarrow b\bar{b}$ decay with the ATLAS detector*, Tech. Rep. ATLAS-CONF-2017-041, CERN, Geneva, Jul, 2017.

- [29] CMS Collaboration, *Evidence for the Higgs boson decay to a bottom quark-antiquark pair*, [arXiv:1709.0749](#).
- [30] ATLAS Collaboration, *Evidence for the spin-0 nature of the Higgs boson using ATLAS data*, *Phys. Lett. B* **726** (2013) 120–144, [[arXiv:1307.1432](#)].
- [31] CMS Collaboration, *Study of the Mass and Spin-Parity of the Higgs Boson Candidate Via Its Decays to Z Boson Pairs*, *Phys. Rev. Lett.* **110** (2013), no. 8 081803, [[arXiv:1212.6639](#)].
- [32] CDF Collaboration, *Observation of top quark production in $\bar{p}p$ collisions*, *Phys. Rev. Lett.* **74** (1995) 2626–2631, [[hep-ex/9503002](#)].
- [33] D0 Collaboration, *Observation of the top quark*, *Phys. Rev. Lett.* **74** (1995) 2632–2637, [[hep-ex/9503003](#)].
- [34] M. Cacciari, M. Czakon, M. Mangano, A. Mitov, and P. Nason, *Top-pair production at hadron colliders with next-to-next-to-leading logarithmic soft-gluon resummation*, *Phys. Lett. B* **710** (2012) 612, [[arXiv:1111.5869](#)].
- [35] P. Bärnreuther, M. Czakon, and A. Mitov, *Percent Level Precision Physics at the Tevatron: First Genuine NNLO QCD Corrections to $q\bar{q} \rightarrow t\bar{t} + X$* , *Phys. Rev. Lett.* **109** (2012) 132001, [[arXiv:1204.5201](#)].
- [36] M. Czakon and A. Mitov, *NNLO corrections to top-pair production at hadron colliders: the all-fermionic scattering channels*, *JHEP* **1212** (2012) 054, [[arXiv:1207.0236](#)].
- [37] M. Czakon and A. Mitov, *NNLO corrections to top pair production at hadron colliders: the quark-gluon reaction*, *JHEP* **1301** (2013) 080, [[arXiv:1210.6832](#)].
- [38] M. Czakon, P. Fiedler, and A. Mitov, *Total Top-Quark Pair-Production Cross Section at Hadron Colliders Through $O(\alpha_S^4)$* , *Phys. Rev. Lett.* **110** (2013) 252004, [[arXiv:1303.6254](#)].
- [39] M. Czakon and A. Mitov, *Top++: A Program for the Calculation of the Top-Pair Cross-Section at Hadron Colliders*, *Comput. Phys. Commun.* **185** (2014) 2930, [[arXiv:1112.5675](#)].
- [40] LHCTopWG-LHC Top Physics Working Group, 2016. <http://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCTopWG/>.
- [41] S. Höche, *Introduction to parton-shower event generators*, in *Proceedings, Theoretical Advanced Study Institute in Elementary Particle Physics: Journeys Through the Precision Frontier: Amplitudes for Colliders (TASI 2014): Boulder, Colorado, June 2-27, 2014*, pp. 235–295, 2015. [arXiv:1411.4085](#).
- [42] K. G. Wilson, *Confinement of Quarks*, *Phys. Rev.* **D10** (1974) 2445–2459.

- [43] D. J. Gross and F. Wilczek, *Ultraviolet Behavior of Nonabelian Gauge Theories*, *Phys. Rev. Lett.* **30** (1973) 1343–1346.
- [44] I.J.R. Aitchison and A.J.G. Hey, *Gauge Theories in Particle Physics, 3rd Edition (2 Volume Set) (Graduate Student Series in Physics)*. Taylor & Francis, 2004.
- [45] J. C. Collins, D. E. Soper, and G. F. Sterman, *Factorization of Hard Processes in QCD*, *Adv. Ser. Direct. High Energy Phys.* **5** (1989) 1–91, [hep-ph/0409313].
- [46] R. Placakyte, *Parton Distribution Functions*, in *Proceedings, 31st International Conference on Physics in collisions (PIC 2011): Vancouver, Canada, August 28-September 1, 2011*, 2011. [arXiv:1111.5452](#).
- [47] ZEUS and H1 Collaborations, *Measurements of deep inelastic scattering at HERA*, in *Proceedings, 32nd International Symposium on Physics in Collision (PIC 2012): Strbske Pleso, Slovakia, September 12-15, 2012*, pp. 93–106, 2013. [arXiv:1301.7572](#).
- [48] G. Altarelli and G. Parisi, *Asymptotic Freedom in Parton Language*, *Nucl. Phys.* **B126** (1977) 298–318.
- [49] V. N. Gribov and L. N. Lipatov, *Deep inelastic $e p$ scattering in perturbation theory*, *Sov. J. Nucl. Phys.* **15** (1972) 438–450.
- [50] Y. L. Dokshitzer, *Calculation of the Structure Functions for Deep Inelastic Scattering and $e^+ e^-$ Annihilation by Perturbation Theory in Quantum Chromodynamics.*, *Sov. Phys. JETP* **46** (1977) 641–653.
- [51] G. Altarelli, *QCD evolution equations for parton densities*, *Scholarpedia* **4** (2009), no. 1 7124. revision #91681.
- [52] H.-L. Lai, M. Guzzi, J. Huston, Z. Li, P. M. Nadolsky, J. Pumplin, and C. P. Yuan, *New parton distributions for collider physics*, *Phys. Rev.* **D82** (2010) 074024, [[arXiv:1007.2241](#)].
- [53] R. D. Ball et al., *Parton distributions with LHC data*, *Nucl. Phys.* **B867** (2013) 244–289, [[arXiv:1207.1303](#)].
- [54] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt, *Parton distributions for the LHC*, *Eur. Phys. J.* **C63** (2009) 189–285, [[arXiv:0901.0002](#)].
- [55] R. D. Ball et al., *Parton distributions with LHC data*, *Nucl. Phys.* **B867** (2013) 244–289, [[arXiv:1207.1303](#)].
- [56] S. Carrazza, S. Forte, and J. Rojo, *Parton Distributions and Event Generators*, in *Proceedings, 43rd International Symposium on Multiparticle Dynamics (ISMD 13)*, pp. 89–96, 2013. [arXiv:1311.5887](#).

- [57] J. Montejo Berlingen and A. Juste Rozas, *Search for new physics in $t\bar{t}$ final states with additional heavy-flavor jets with the ATLAS detector*. PhD thesis, IFAE, Apr, 2016. CERN-THESIS-2015-140.
- [58] B. R. Webber, *A QCD Model for Jet Fragmentation Including Soft Gluon Interference*, *Nucl. Phys.* **B238** (1984) 492–528.
- [59] G. Marchesini and B. R. Webber, *Monte Carlo Simulation of General Hard Processes with Coherent QCD Radiation*, *Nucl. Phys.* **B310** (1988) 461–526.
- [60] B. Andersson, G. Gustafson, G. Ingelman, and T. Sjostrand, *Parton Fragmentation and String Dynamics*, *Phys. Rept.* **97** (1983) 31–145.
- [61] T. Sjostrand, *Jet Fragmentation of Nearby Partons*, *Nucl. Phys.* **B248** (1984) 469–502.
- [62] D. Amati and G. Veneziano, *Preconfinement as a property of perturbative QCD*, *Physics Letters B* **83** (Apr., 1979) 87–92.
- [63] B. R. Webber, *Hadronization*, in *Proceedings: Summer School on Hadronic Aspects of Collider Physics, Zuoz, Switzerland, Aug 23-31, 1994*, pp. 49–77, 1994. [hep-ph/9411384](#).
- [64] ATLAS Collaboration, *Measurement of the underlying event in jet events from 7 TeV proton–proton collisions with the ATLAS detector*, *Eur. Phys. J. C* **74** (2014) 2965, [[arXiv:1406.0392](#)].
- [65] P. Nason, *A New method for combining NLO QCD with shower Monte Carlo algorithms*, *JHEP* **0411** (2004) 040, [[hep-ph/0409146](#)].
- [66] E. Re, *Single-top Wt -channel production matched with parton showers using the POWHEG method*, *Eur. Phys. J. C* **71** (2011) 1547, [[arXiv:1009.2450](#)].
- [67] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, *JHEP* **07** (2014) 079, [[arXiv:1405.0301](#)].
- [68] S. Frixione and B. R. Webber, *Matching NLO QCD computations and parton shower simulations*, *JHEP* **06** (2002) 029, [[hep-ph/0204244](#)].
- [69] T. Sjostrand, S. Mrenna, and P. Z. Skands, *A Brief Introduction to PYTHIA 8.1*, *Comput. Phys. Commun.* **178** (2008) 852–867, [[arXiv:0710.3820](#)].
- [70] M. Bahr et al., *Herwig++ physics and manual*, *Eur. Phys. J. C* **58** (2008) 639–707, [[arXiv:0803.0883](#)].

- [71] J. Bellm et al., *Herwig 7.0/Herwig++ 3.0 release note*, *Eur. Phys. J.* **C76** (2016), no. 4 196, [[arXiv:1512.0117](#)].
- [72] D. J. Lange, *The EvtGen particle decay simulation package*, *Nucl. Instrum. Meth. A* **462** (2001) 152–155.
- [73] T. Gleisberg, S. Höche, F. Krauss, M. Schönherr, S. Schumann, et al., *Event generation with SHERPA 1.1*, *JHEP* **0902** (2009) 007, [[arXiv:0811.4622](#)].
- [74] S. Schumann and F. Krauss, *A Parton shower algorithm based on Catani-Seymour dipole factorisation*, *JHEP* **0803** (2008) 038, [[arXiv:0709.1027](#)].
- [75] F. Cascioli, P. Maierhofer, and S. Pozzorini, *Scattering Amplitudes with Open Loops*, *Phys. Rev. Lett.* **108** (2012) 111601, [[arXiv:1111.5206](#)].
- [76] GEANT4 Collaboration, *GEANT4: A Simulation toolkit*, *Nucl. Instrum. Meth.* **A506** (2003) 250–303.
- [77] ATLAS Collaboration, *Fast Simulation for ATLAS: Atlfast-II and ISF*, *J. Phys. Conf. Ser.* **396** (2012) 022031.
- [78] L. Evans and P. Bryant, *LHC Machine*, *JINST* **3** (2008) S08001.
- [79] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, *JINST* **3** (2008) S08003.
- [80] CMS Collaboration, *The CMS experiment at the CERN LHC*, *JINST* **3** (2008) S08004.
- [81] F. Marcastel, *CERN’s Accelerator Complex. La chaîne des accélérateurs du CERN*, Oct, 2013. General Photo, <https://cds.cern.ch/record/1621583>.
- [82] T. Kawamoto et al., *New Small Wheel Technical Design Report*, Tech. Rep. CERN-LHCC-2013-006. ATLAS-TDR-020, Jun, 2013. ATLAS New Small Wheel Technical Design Report.
- [83] M. Aleksa et al., *ATLAS Liquid Argon Calorimeter Phase-I Upgrade Technical Design Report*, Tech. Rep. CERN-LHCC-2013-017. ATLAS-TDR-022, Sep, 2013. Final version presented to December 2013 LHCC.
- [84] L. Drosdal, *LHC Injection Beam Quality During LHC Run I*. PhD thesis, CERN, 2015-03-03. CERN-THESIS-2015-254.
- [85] R. Bruce et al., *LHC Run 2: Results and Challenges*, Tech. Rep. CERN-ACC-2016-0103, CERN, Geneva, Jul, 2016.
- [86] ATLAS Luminosity Group, *Luminosity public results run2*, July, 2015. <https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2>.

- [87] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, *JINST* **3** (2008) S08003.
- [88] ATLAS Collaboration, *Track Reconstruction Performance of the ATLAS Inner Detector at $\sqrt{s} = 13$ TeV*, Tech. Rep. ATL-PHYS-PUB-2015-018, CERN, Geneva, Jul, 2015.
- [89] ATLAS Collaboration, *Alignment of the ATLAS Inner Detector and its Performance in 2012*, Tech. Rep. ATLAS-CONF-2014-047, CERN, Geneva, Jul, 2014.
- [90] M. Capeans et al., *ATLAS Insertable B-Layer Technical Design Report*, Tech. Rep. CERN-LHCC-2010-013. ATLAS-TDR-19, CERN, Geneva, 2010.
- [91] ATLAS Collaboration, *Expected performance of the ATLAS b-tagging algorithms in Run-2*, Tech. Rep. ATL-PHYS-PUB-2015-022, CERN, Geneva, Jul, 2015.
- [92] ATLAS Collaboration, *ATLAS: Detector and physics performance technical design report. Volume 1*, Tech. Rep. CERN-LHCC-99-14, ATLAS-TDR-14, 1999.
- [93] G. F. Knoll, *Radiation detection and measurement; 4th ed.* Wiley, New York, NY, 2010.
- [94] P. Adragna et al., *Testbeam studies of production modules of the ATLAS tile calorimeter*, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **606** (2009), no. 3 362 – 394.
- [95] F. Bauer et al., *Construction and test of MDT chambers for the ATLAS muon spectrometer*, *Nucl. Instrum. Meth.* **A461** (2001) 17–20.
- [96] T. Argyropoulos et al., *Cathode strip chambers in ATLAS: Installation, commissioning and in situ performance*, *IEEE Trans. Nucl. Sci.* **56** (2009) 1568–1574.
- [97] G. Aielli et al., *The RPC first level muon trigger in the barrel of the ATLAS experiment*, *Nucl. Phys. Proc. Suppl.* **158** (2006) 11–15.
- [98] S. Majewski, G. Charpak, A. Breskin, and G. Mikenberg, *A thin multiwire chamber operating in the high multiplication mode*, *Nucl. Instrum. Meth.* **217** (1983) 265–271.
- [99] S. Artz et al., *Upgrade of the ATLAS central trigger for LHC run-2*, *Journal of Instrumentation* **10** (2015), no. 02 C02030.
- [100] ATLAS Collaboration, *ATLAS Computing: technical design report*. Technical Design Report ATLAS. CERN, Geneva, 2005.

- [101] M. Shochet, L. Tompkins, V. Cavaliere, P. Giannetti, A. Annovi, and G. Volpi, *Fast TracKer (FTK) Technical Design Report*, Tech. Rep. CERN-LHCC-2013-007. ATLAS-TDR-021, Jun, 2013.
- [102] N. Asbah, *A hardware fast tracker for the ATLAS trigger*, *Phys. Part. Nucl. Lett.* **13** (2016), no. 5 527–531.
- [103] ATLAS Collaboration, *Luminosity determination in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector at the LHC*, *Eur. Phys. J.* **C76** (2016), no. 12 653, [[arXiv:1608.0395](#)].
- [104] V. Cindro et al., *The ATLAS beam conditions monitor*, *JINST* **3** (2008) P02004.
- [105] P. Jenni, M. Nordberg, M. Nessi, and K. Jon-And, *ATLAS Forward Detectors for Measurement of Elastic Scattering and Luminosity*. Technical Design Report ATLAS. CERN, Geneva, 2008.
- [106] R. G. Newton, *Optical theorem and beyond*, *American Journal of Physics* **44** (July, 1976) 639–642.
- [107] S. Jakobsen, *Commissioning of the Absolute Luminosity For ATLAS detector at the LHC*. PhD thesis, CERN, 2013-12-16. CERN-THESIS-2013-230.
- [108] A. Sopczak et al., *MPX detectors as LHC luminosity monitor*, in *Proceedings, 2015 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC 2015): San Diego, California, United States*, p. 7581870, 2016.
- [109] S. van der Meer, *Calibration of the Effective Beam Height in the ISR*, .
- [110] T. Cornelissen, M. Elsing, S. Fleischmann, W. Liebig, E. Moyse, and A. Salzburger, *Concepts, Design and Implementation of the ATLAS New Tracking (NEWT)*, Tech. Rep. ATL-SOFT-PUB-2007-007, CERN, Geneva, 2007.
- [111] ATLAS Collaboration, *The Optimization of ATLAS Track Reconstruction in Dense Environments*, Tech. Rep. ATL-PHYS-PUB-2015-006, CERN, Geneva, 2015.
- [112] ATLAS Collaboration, *Reconstruction of primary vertices at the ATLAS experiment in Run 1 proton-proton collisions at the LHC*, *Eur. Phys. J.* **C77** (2017), no. 5 332, [[arXiv:1611.1023](#)].
- [113] ATLAS Collaboration, *Performance of primary vertex reconstruction in proton-proton collisions at $\sqrt{s} = 7$ TeV in the ATLAS experiment*, *ATLAS-CONF-2010-069* (2010).
- [114] ATLAS Collaboration, *Electron reconstruction and identification efficiency measurements with the ATLAS detector using the 2011 LHC proton-proton collision data*, [arXiv:1404.2240](#).

- [115] ATLAS Collaboration, *Electron and photon energy calibration with the ATLAS detector using LHC Run 1 data*, *Eur. Phys. J. C* **74** (2014) 3071, [arXiv:1407.5063].
- [116] ATLAS Collaboration, *Measurements of the photon identification efficiency with the ATLAS detector using 4.9fb^{-1} $p\bar{p}$ collision data collected in 2011*, *ATLAS-CONF-2012-123* (2012).
- [117] ATLAS Collaboration, *Electron efficiency measurements with the ATLAS detector using the 2015 LHC proton-proton collision data*, Tech. Rep. ATLAS-CONF-2016-024, CERN, Geneva, Jun, 2016.
- [118] *Muon Combined Performance in Run 2 (25 ns runs)*, Tech. Rep. ATL-COM-MUON-2015-093, Geneva, Nov, 2015.
- [119] W. Lampl et al., “Calorimeter Clustering Algorithms: Description and Performance.” ATL-LARG-PUB-2008-002, 2008.
- [120] ATLAS Collaboration, *Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1*, Tech. Rep. CERN-PH-EP-2015-304, CERN, Geneva, Mar, 2016.
- [121] G. P. Salam, *Towards Jetography*, *Eur. Phys. J. C* **67** (2010) 637, [arXiv:0906.1833].
- [122] M. Cacciari, G. P. Salam, and G. Soyez, *The Anti- $k(t)$ jet clustering algorithm*, *JHEP* **04** (2008) 063, [arXiv:0802.1189].
- [123] ATLAS Collaboration, *Jet Calibration and Systematic Uncertainties for Jets Reconstructed in the ATLAS Detector at $\sqrt{s} = 13\text{ TeV}$* , Tech. Rep. ATL-PHYS-PUB-2015-015, CERN, Geneva, Jul, 2015.
- [124] ATLAS Collaboration, *Pile-up subtraction and suppression for jets in ATLAS*, Tech. Rep. ATLAS-CONF-2013-083, CERN, Geneva, Aug, 2013.
- [125] M. Cacciari, G. P. Salam, and G. Soyez, *The Catchment Area of Jets*, *JHEP* **0804** (2008) 005.
- [126] ATLAS Collaboration, *Data-driven determination of the energy scale and resolution of jets reconstructed in the ATLAS calorimeters using dijet and multijet events at $\sqrt{s} = 8\text{ TeV}$* , Tech. Rep. ATLAS-CONF-2015-017, CERN, Geneva, Apr, 2015.
- [127] ATLAS Collaboration, *Determination of the jet energy scale and resolution at ATLAS using Z/γ -jet events in data at $\sqrt{s} = 8\text{ TeV}$* , Tech. Rep. ATLAS-CONF-2015-057, CERN, Geneva, Oct, 2015.

- [128] ATLAS Collaboration, *Jet energy scale measurements and their systematic uncertainties in proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, arXiv:1703.0966.
- [129] ATLAS Collaboration, *Monte Carlo Calibration and Combination of In-situ Measurements of Jet Energy Scale, Jet Energy Resolution and Jet Mass in ATLAS*, Tech. Rep. ATLAS-CONF-2015-037, CERN, Geneva, Aug, 2015.
- [130] ATLAS Collaboration, *Tagging and suppression of pileup jets with the ATLAS detector*, Tech. Rep. ATLAS-CONF-2014-018, CERN, Geneva, May, 2014.
- [131] ATLAS Collaboration, *Commissioning of the ATLAS high-performance b-tagging algorithms in the $\sqrt{s} = 7$ TeV collision data*, ATLAS-CONF-2011-102 (2011).
- [132] ATLAS Collaboration, *A new inclusive secondary vertex algorithm for b-jet tagging in ATLAS*, *J. Phys. Conf. Ser.* **119** 032032 (2008).
- [133] R. Fruhwirth, *Application of Kalman filtering to track and vertex fitting*, *Nucl. Instrum. Meth. A* **262** (1987) 444.
- [134] ATLAS Collaboration, *Optimisation of the ATLAS b-tagging performance for the 2016 LHC Run*, Tech. Rep. ATL-PHYS-PUB-2016-012, CERN, Geneva, Jun, 2016.
- [135] ATLAS Collaboration, “Search for the Standard Model Higgs boson produced in association with top quarks and decaying into a bb pair in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector.” ATLAS-CONF-2016-080 (2016).
- [136] ATLAS Collaboration, “Calibration of b-tagging using dileptonic top pair events in a combinatorial likelihood approach with the ATLAS experiment.” ATLAS-CONF-2014-004, 2014.
- [137] ATLAS Collaboration, *Measurement of b-tagging Efficiency of c-jets in $t\bar{t}$ Events Using a Likelihood Approach with the ATLAS Detector*, Tech. Rep. ATLAS-CONF-2018-001, CERN, Geneva, Mar, 2018.
- [138] ATLAS Collaboration, *Measurement of the Mistag Rate of b-tagging algorithms with 5 fb^{-1} of Data Collected by the ATLAS Detector*, ATLAS-CONF-2012-040 (2012).
- [139] ATLAS Collaboration, *Expected performance of missing transverse momentum reconstruction for the ATLAS detector at $\sqrt{s} = 13$ TeV*, ATL-PHYS-PUB-2015-023 (2015).
- [140] ALEPH Collaboration, DELPHI Collaboration, L3 Collaboration, OPAL Collaboration, SLD Collaboration, LEP Electroweak Working Group, SLD Electroweak Group, SLD Heavy Flavour Group Collaboration, *Precision electroweak measurements on the Z resonance*, *Phys.Rept.* **427** (206) 257–454.

- [141] ATLAS Collaboration, *Measurement of the isolated di-photon cross-section in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector*, *Phys. Rev.* **D85** (2012) 012003, [[arXiv:1107.0581](#)].
- [142] E. Ritsch, *Fast Calorimeter Punch-Through Simulation for the ATLAS Experiment*. PhD thesis, CERN, 2011. CERN-THESIS-2011-112.
- [143] D0 Collaboration, *Measurement of the $t\bar{t}$ production cross section in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV using kinematic characteristics of lepton + jets events*, *Phys.Rev. D* **76:092007** (Nov 2007).
- [144] ATLAS Collaboration, *Search for Supersymmetry Using Final States with One Lepton, Jets, and Missing Transverse Momentum with the ATLAS Detector in $\sqrt{s} = 7$ TeV pp Collisions*, *Phys. Rev. Lett.* **106** (2011) 131802, [[arXiv:1102.2357](#)].
- [145] ATLAS Collaboration, *Search for an excess of events with an identical flavour lepton pair and significant missing transverse momentum in $\sqrt{s} = 7$ TeV proton–proton collisions with the ATLAS detector*, *Eur. Phys. J. C* **71** (2011) 1647, [[arXiv:1103.6208](#)].
- [146] ATLAS Collaboration, *Observation of spin correlation in $t\bar{t}$ events from pp collisions at $\sqrt{s} = 7$ TeV using the ATLAS detector*, *Phys. Rev. Lett.* **108** (2012) 212001, [[arXiv:1203.4081](#)].
- [147] ATLAS Collaboration, *A search for $t\bar{t}$ resonances with the ATLAS detector in 2.05 fb^{-1} of proton–proton collisions at $\sqrt{s} = 7$ TeV*, *Eur. Phys. J. C* **72** (2012) 2083, [[arXiv:1205.5371](#)].
- [148] ATLAS Collaboration, *A search for $t\bar{t}$ resonances in lepton+jets events with highly boosted top quarks collected in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector*, *JHEP* **09** (2012) 041, [[arXiv:1207.2409](#)].
- [149] ATLAS Collaboration, *Search for resonant top plus jet production in $t\bar{t} + \text{jets}$ events with the ATLAS detector in pp collisions at $\sqrt{s} = 7$ TeV*, *Phys. Rev. D* **86** (2012) 091103, [[arXiv:1209.6593](#)].
- [150] ATLAS Collaboration, *Search for the Standard Model Higgs boson produced in association with top quarks and decaying into a $b\bar{b}$ pair in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, [arXiv:1712.0889](#).
- [151] ATLAS Collaboration, *Measurements of fiducial cross-sections for $t\bar{t}$ production with one or two additional b-jets in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector*, *Eur. Phys. J. C* **76** (2016) 11, [[arXiv:1508.0686](#)].
- [152] ATLAS Collaboration, *Search for the Standard Model Higgs boson produced in association with top quarks and decaying into $b\bar{b}$ in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, *Eur. Phys. J. C* **75** (2015) 349, [[arXiv:1503.0506](#)].

- [153] ATLAS Collaboration, *Search for the Standard Model Higgs boson decaying into $b\bar{b}$ produced in association with top quarks decaying hadronically in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, *JHEP* **05** (2016) 160, [[arXiv:1604.0381](#)].
- [154] CMS Collaboration, *Search for the associated production of the Higgs boson with a top-quark pair*, *JHEP* **09** (2014) 087, [[arXiv:1408.1682](#)].
- [155] ATLAS Collaboration, *Luminosity determination in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector at the LHC*, [arXiv:1608.0395](#).
- [156] ATLAS Collaboration, “2015 start-up trigger menu and initial performance assessment of the ATLAS trigger using Run-2 data.” ATL-DAQ-PUB-2016-001, 2016.
- [157] ATLAS Collaboration, *Atlas pythia 8 tunes to 7 tev data*, *ATL-PHYS-PUB-2014-021* (2014). <https://cds.cern.ch/record/1966419>.
- [158] NNPDF Collaboration, *Parton distributions for the LHC Run II*, *JHEP* **04** (2015) 040, [[arXiv:1410.8849](#)].
- [159] P. Artoisenet, R. Frederix, O. Mattelaer, and R. Rietkerk, *Automatic spin-entangled decays of heavy resonances in Monte Carlo simulations*, *JHEP* **03** (2013) 015, [[arXiv:1212.3460](#)].
- [160] R. Raitio and W. W. Wada, *Higgs Boson Production at Large Transverse Momentum in QCD*, *Phys. Rev.* **D19** (1979) 941.
- [161] W. Beenakker et al. *Nucl. Phys.* **B653** (2003) 151–203, [[hep-ph/0211352](#)].
- [162] S. Dawson, C. Jackson, L. H. Orr, L. Reina, and D. Wackeroth *Phys. Rev.* **D68** (2003) 034022, [[hep-ph/0305087](#)].
- [163] Y. Zhang, W.-G. Ma, R.-Y. Zhang, C. Chen, and L. Guo, *QCD NLO and EW NLO corrections to $t\bar{t}H$ production with top quark decays at hadron collider*, *Phys. Lett.* **B738** (2014) 1–5, [[arXiv:1407.1110](#)].
- [164] S. Frixione, V. Hirschi, D. Pagani, H.-S. Shao, and M. Zaro, *Electroweak and QCD corrections to top-pair hadroproduction in association with heavy bosons*, *JHEP* **06** (2015) 184, [[arXiv:1504.0344](#)].
- [165] ATLAS Collaboration, “Further studies on simulation of top-quark production for the ATLAS experiment at $\sqrt{s} = 13$ TeV.” ATL-PHYS-PUB-2016-016, 2016.
- [166] S. Frixione, P. Nason, and C. Oleari, *Matching NLO QCD computations with Parton Shower simulations: the POWHEG method*, *JHEP* **0711** (2007) 070, [[arXiv:0709.2092](#)].

- [167] S. Alioli, P. Nason, C. Oleari, and E. Re, *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*, *JHEP* **1006** (2010) 043, [[arXiv:1002.2581](#)].
- [168] J. M. Campbell, R. K. Ellis, P. Nason, and E. Re, *Top-pair production and decay at NLO matched with parton showers*, *JHEP* **04** (2015) 114, [[arXiv:1412.1828](#)].
- [169] ATLAS Collaboration, *ATLAS Run 1 Pythia8 tunes*, Tech. Rep. ATL-PHYS-PUB-2014-021, CERN, Geneva, Nov, 2014.
- [170] ATLAS Collaboration, “Studies on top-quark Monte Carlo modelling for Top2016.” ATL-PHYS-PUB-2016-020, 2016.
- [171] ATLAS Collaboration, “Studies on top-quark Monte Carlo modelling with Sherpa and MG5_aMC@NLO.” ATL-PHYS-PUB-2017-007, 2017.
- [172] ATLAS Collaboration, “Measurements of top-quark pair differential cross-sections in the lepton+jets channel in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector.” ATLAS-CONF-2016-040, 2016.
- [173] F. Cascioli, P. Maierhofer, N. Moretti, S. Pozzorini, and F. Siegert, *NLO matching for $t\bar{t}b\bar{b}$ production with massive b -quarks*, *Phys. Lett. B* **734** (2014) 210, [[arXiv:1309.5912](#)].
- [174] S. Alioli, P. Nason, C. Oleari, and E. Re, *NLO single-top production matched with shower in POWHEG: s - and t -channel contributions*, *JHEP* **09** (2009) 111, [[arXiv:0907.4076](#)].
- [175] T. Gleisberg and S. Höche, *Comix, a new matrix element generator*, *JHEP* **0812** (2008) 039, [[arXiv:0808.3674](#)].
- [176] S. Höche, F. Krauss, M. Schönherr, and F. Siegert, *QCD matrix elements + parton showers: The NLO case*, *JHEP* **04** (2013) 027, [[arXiv:1207.5030](#)].
- [177] J. Butterworth et al., *Single Boson and Diboson Production Cross Sections in pp Collisions at $\sqrt{s}=7$ TeV*, ATL-COM-PHYS-2010-695 (2010). <https://cds.cern.ch/record/1287902>.
- [178] S. Frixione, E. Laenen, P. Motylinski, B. R. Webber, and C. D. White, *Single-top hadroproduction in association with a W boson*, *JHEP* **0807** (2008) 029, [[arXiv:0805.3067](#)].
- [179] N. Kidonakis, *Two-loop soft anomalous dimensions for single top quark associated production with a W - or H -*, *Phys. Rev. D* **82** (2010) 054018, [[arXiv:1005.4451](#)].
- [180] N. Kidonakis, *NNLL resummation for s -channel single top quark production*, *Phys. Rev. D* **81** (2010) 054028, [[arXiv:1001.5034](#)].

- [181] N. Kidonakis, *Next-to-next-to-leading-order collinear and soft gluon corrections for t -channel single top quark production*, *Phys. Rev. D* **83** (2011) 091503, [[arXiv:1103.2792](#)].
- [182] ATLAS Collaboration, *The ATLAS Simulation Infrastructure*, *Eur. Phys. J. C* **70** (2010) 823, [[arXiv:1005.4568](#)].
- [183] ATLAS Collaboration, “The simulation principle and performance of the ATLAS fast calorimeter simulation FastCaloSim.” ATL-PHYS-PUB-2010-013, 2010.
- [184] ATLAS Collaboration, *Evidence for the associated production of the Higgs boson and a top quark pair with the ATLAS detector*, [arXiv:1712.0889](#).
- [185] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.
- [186] Y. Coadou, “Boosted decision trees and applications.” EPJ Web of conferences 55 (2013).
- [187] A. Hocker et al., *TMVA - Toolkit for Multivariate Data Analysis*, *PoS ACAT* (2007) 040, [[physics/0703039](#)].
- [188] Y. Freund and R. E. Schapire, *Experiments with a new boosting algorithm*, in *IN PROCEEDINGS OF THE THIRTEENTH INTERNATIONAL CONFERENCE ON MACHINE LEARNING*, pp. 148–156, Morgan Kaufmann, 1996.
- [189] A. Bredenstein, A. Denner, S. Dittmaier, and S. Pozzorini, *NLO QCD Corrections to Top Anti-Top Bottom Anti-Bottom Production at the LHC: 2. full hadronic results*, *JHEP* **1003** (2010) 021, [[arXiv:1001.4006](#)].
- [190] G. Bevilacqua, M. Czakon, C. Papadopoulos, R. Pittau, and M. Worek, *Assault on the NLO Wishlist: $pp \rightarrow t\bar{t}b\bar{b}$* , *JHEP* **0909** (2009) 109, [[arXiv:0907.4723](#)].
- [191] A. Bredenstein, A. Denner, S. Dittmaier, and S. Pozzorini, *NLO QCD corrections to $pp \rightarrow t\bar{t}b\bar{b} + X$ at the LHC*, *Phys. Rev. Lett.* **103** (2009) 012002, [[arXiv:0905.0110](#)].
- [192] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt, *Parton distributions for the LHC*, *Eur. Phys. J. C* **63** (2009) 189–285, [[arXiv:0901.0002](#)].
- [193] ATLAS Collaboration, *Performance of b -jet identification in the ATLAS experiment*, *JINST* **11** (2016) P04008, [[arXiv:1512.0109](#)].
- [194] ATLAS Collaboration, “Studies of $t\bar{t}c\bar{c}$ production with MADGRAPH5_AMC@NLO and HERWIG++ for the ATLAS experiment.” ATL-PHYS-PUB-2016-011, 2016.
- [195] ATLAS Collaboration, “Multi-boson simulation for 13 TeV ATLAS analyses.” ATL-PHYS-PUB-2016-002, 2016.

- [196] J. M. Campbell and R. K. Ellis, $t\bar{t}W^{+-}$ production and decay at NLO, *JHEP* **07** (2012) 052, [[arXiv:1204.5678](#)].
- [197] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, *Eur. Phys. J.* **C71** (2011) 1554, [[arXiv:1007.1727](#)]. [Erratum: *Eur. Phys. J.* **C73**,2501(2013)].
- [198] ROOT Collaboration, *HistFactory: A tool for creating statistical models for use with RooFit and RooStats*, Tech. Rep. CERN-OPEN-2012-016, 2012.
- [199] W. Verkerke and D. P. Kirkby, *The RooFit toolkit for data modeling*, *eConf C0303241* (2003) MOLT007, [[physics/0306116](#)].
- [200] W. Verkerke and D. Kirkby, “Roofit users manual.” <http://roofit.sourceforge.net/>.
- [201] R. M. L. Team, “Minuit2 minimization package.” <http://project-mathlibs.web.cern.ch/project-mathlibs/sw/Minuit2/html/index.html>.
- [202] G. Cowan, *Statistical Data Analysis*. Oxford science publications. Clarendon Press, 1998.
- [203] A. L. Read, *Presentation of search results: The CL_S technique*, *J. Phys. G* **28** (2002) 2693.
- [204] T. Junk, *Confidence level computation for combining searches with small statistics*, *Nucl. Instrum. Meth. A* **434** (1999) 435, [[hep-ex/9902006](#)].
- [205] B. Nachman, P. Nef, A. Schwartzman, M. Swiatlowski, and C. Wanotayaroj, *Jets from Jets: Re-clustering as a tool for large radius jet reconstruction and grooming at the LHC*, *JHEP* **02** (2015) 075, [[arXiv:1407.2922](#)].